WORKING PAPER SERIES

Forecasting financial markets with semantic network analysis in the COVID—19 crisis

Andrea Fronzetti Colladon, Stefano Grassi,
Francesco Ravazzolo, Francesco Violante

N°06 / Feb 2021

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

Forecasting financial markets with semantic network analysis in the COVID-19 crisis*

Andrea Fronzetti Colladon[†] Stefano Grassi[‡]
Francesco Ravazzolo[§] Francesco Violante[¶]

[†]University of Perugia
[‡]University of Rome 'Tor Vergata'
[§]Free University of Bozen-Bolzano and CAMP, BI Norwegian Business School
[¶]ENSAE - École nationale de la Statistique et de l' Administration Économique, CREST and CREATES

November 30, 2020

Abstract

This paper uses a new textual data index for predicting stock market data. The index is applied to a large set of news to evaluate the importance of one or more general economic related keywords appearing in the text. The index assesses the importance of the economic related keywords, based on their frequency of use and semantic network position. We apply it to the Italian press and construct indices to predict Italian stock and bond market returns and volatilities in a recent sample period, including the COVID-19 crisis. The evidence shows that the index captures the different phases of financial time series well. Moreover, results indicate strong evidence of predictability for bond market data, both returns and volatilities, short and long maturities, and stock market volatility.

*The authors are grateful to Vincenzo D'Innella Capano, CEO of Telpress International B.V., and to Lamberto Celommi, for making the news data available and for the support received during the data collection process. The computing resources and the related technical support used for this study were provided by CRESCO/ENEAGRID High Performance Computing infrastructure and its staff. CRESCO/ENEAGRID High Performance Computing infrastructure is funded by ENEA, the Italian National Agency for New Technologies, Energy and Sustainable Economic Development and by Italian and European research programs, see http://www.cresco.enea.it/english for information. This paper is part of the research activities at the Centre for Applied Macroeconomics and Commodity Prices (CAMP) at the BI Norwegian Business School. This study is also partially supported by the University of Perugia, through the program Fondo Ricerca di Base 2019", project no. RICBA19LT ("Business Intelligence, Data Analytics e simulazioni a eventi discreti per le Smart Companies nell'era dell'Industria 4.0"). Stefano Grassi gratefully acknowledges financial support from the University of Rome 'Tor Vergata' under the "Beyond Borders" grant (CUP: E84I20000900005). Francesco Violante acknowledges financial support from the grant of the French National Research Agency (ANR), Investissements d'Avenir (LabEx Ecodec/ANR-11-LABX-0047), and of the Danish Council for Independent Research (1329-00011A). The funding organisations had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

1 Introduction

In order to make informed decisions, individuals attempt to anticipate future movements of economic variables and business cycle dynamics. This is particularly important for investment decisions in financial markets. An agent can gain from predicting future market needs and investing early, satisfying market demands when they appear. There is a large body of literature on predicting financial markets and many indicators have been used as predictors. However, the predictability power of such indicators is often questionable and evidence of systematic predictability is weak, see Welch and Goyal (2008).

More recently, a new theory based on a different type of data has received increasing attention, namely the news media perception and evaluation of the business cycles, see Beaudry and Portier (2006). Beaudry and Portier (2014) state that "according to the news media view of the business cycle, both the boom and the bust are direct consequences of people's incentive to speculate on information related to future developments of the economy." One of the main challenges of this theory is the definition of news and how to empirically measure it. To overcome the problem, Larsen and Thorsrud (2019) propose a novel and direct measure of media news based on their semantic content. Using text data from a Norwegian financial newspaper, they document a superior predictability power of their indicator for Norwegian stock indices returns.

The recent COVID-19 crisis has rendered the task of predicting financial market fluctuations even more difficult. Markets have promptly reacted to early news related to the coronavirus. Indeed, as early as March 2020, markets collapsed, with weekly losses above 30%, well before macro-economic indicators were impacted, as well as accurate information about the severity of the spread of the coronavirus in Europe and in the US was available. In this context, Italy represents a peculiar case. Indeed, it was the first country in Europe to experience a major outbreak of COVID-19, with a much higher mortality rate than observed in other countries. This crisis hit Italy in a moment when public finances were already under stress. In Italy, the country with the third largest public debt in the word after US and Japan, the COVID-19 crisis has intensified an already difficult economic situation, posing serious doubts about the short-term sustainability of the economy, as well as the long-term outlook. From early summer 2020, signals and actions from the European Central Bank and the European Union in support of the members' economies and health systems have somewhat alleviated the burden of an otherwise extensively paralysed economic environment.

Such period of economic and social turmoil, where the news media have not merely covered the role of broadcasting information but also that of conveying perceptions and expectations about future states of the economy, represents an unprecedented testing ground to evaluate the link between news media information and macro-finance variables.

This paper introduces a new index of semantic importance. The index is based on a novel methodology that evaluates the relative importance of one or more general economic related keywords (ERKs) that appear in the news. The index, whose construction combines methods drawn from both network analysis and text mining, evaluates semantic importance along the three dimensions of prevalence, i.e. frequency of word occurrences, connectivity, i.e. degree of centrality of a word in the discourse, and diversity, i.e. richness and distinctiveness of textual associations.

Previous research mainly looked at media sentiment or media coverage, without analysing the embeddedness that ERKs have in the corpus and their relationships with other words. Fronzetti Colladon et al. (2020) suggest that sentiment can be much less informative than semantic importance. This is sometimes attributable to the methodologies used to calculate sentiment, for example when trained on a general domain and then applied to a specific context. Indeed, sentiment algorithms have variable error rates and their reliability has been questioned in different studies (Jussila et al., 2017). Accordingly, our approach excludes sentiment from the main indicator, thus measuring importance, and not favourability, of ERKs.

We identify 38 relevant ERKs suited to our scope. Using a large database of articles published by a pool of Italian newspapers, over the period spanning between January 2017 and August 2020, we assign a score to each ERK, compounding the three dimensions mentioned above. We then aggregate the information from the 38 ERKs in a single composite news index. For the aggregation, we apply Partial Least Squares (PLS) between the target variable and the (38 ERKs) predictors, incorporating information from both the definition of scores and loadings. de Jong (1993), show that the scores and loadings can be chosen so as to maximize the covariance between the dependent variable and the predictors. To do so, our methodology based on PLS allows us to construct a composite index in a target specific manner.

While existing empirical literature in this context has focused on stock market return predictability, see Baker and Wurgler (2006), Baker and Wurgler (2007), Chung et al. (2012) and Limongi Concetto and Ravazzolo (2019) among others, we evaluate the power of our media news index in predicting not only the Italian stock market aggregate return but also various short and long maturity government bonds index returns, as well as their volatility.

Periods of large movements in the stock and bond markets are associated with political instability and economic uncertainty. Our findings show that the index is able to anticipate the different phases of the market and capture idiosyncratic features of each series. We find evidence of economically meaningful and statistically significant predictability for government bond returns and volatilities. For stock market data, we find evidence of predictability of the market portfolio returns only to a limited extent. However, when predicting stock market volatility, adding information contained in the news media improves the prediction accuracy up to 9%, compared to standard benchmark forecast models. Alternative (standard) methods for dealing with newspaper information, such as the sentiment index, do not offer similar gains.

The remainder of the paper is organised as follows. Section 2 introduces our new textual data index. Section 3 provides a detailed description of the data employed, the methodological strategy used to predict financial time series with textual data, and the results of our analysis. Finally Section 4 concludes.

2 A new index for textual data

In this paper we propose a novel measure of semantic importance that combines methods drawn from both social network analysis and text mining. The index aims at measuring the relative importance of a predefined set of words mentioned in a large set of textual documents. The methodology labelled Semantic Brand Score was introduced by Fronzetti Colladon (2018) for application to commercial brands' reputation and awareness, but has never been applied in the economic and financial environment. Starting from the word frequency as a natural measure of importance within a text (Piantadosi, 2014), the association that a word has in the text, as well as the heterogeneity of its context, are used as pivotal additional variables for a comprehensive assessment. Our index explicitly exploits the relationships among words in a text. To this end, texts are transformed into networks of co-occurring words and relationships are studied through social network analysis, see Wasserman and Faust (1994). As an example consider the following sentence (and a word co-occurrence threshold of 3 words) to generate the network reported in Figure 1: "The proud and unfeeling landlord views his extensive fields, and without a thought for the wants of his brethren, in imagination consumes himself the whole harvest" (from The Theory of Moral Sentiments of Adam Smith). Words are presented without stemming, for the sake of readability.

Figure 1: *Graphical representation of the sentence "The proud and unfeeling landlord views his extensive fields, and without a thought for the wants of his brethren, in imagination consumes himself the whole harvest" (from The Theory of Moral Sentiments of Adam Smith)*

The index measures words' semantic importance, i.e. in our context of the selected ERKs, along the three dimensions of prevalence, diversity and connectivity.

Prevalence, which relates to the notion of awareness, see Keller (1993), measures how frequently an ERK is mentioned, the rationality being that an ERK used frequently is easier to remember and more recognisable. Prevalence is calculated as the frequency of a word in a given set of documents and time-frame. Prevalence of a particular set of words could ultimately influence the opinions and behaviours of the readers. For instance, in our context, the recurrent use of specific combinations of words may trigger fear or an optimistic view about the current, as well as future, states of the economy.

Diversity measures the degree of heterogeneity of the semantic context in which a word is used, with emphasis on the richness and distinctiveness of its textual associations. Diversity is defined by the number and uniqueness of connections a word has in the co-occurrence network and it is measured by the distinctiveness centrality metric introduced in Fronzetti Colladon and Naldi

(2020). More precisely, in a graph of n nodes (words) and E edges (e.g. 1), distinctiveness of node i is given by:

$$D_i = \sum_{j=1, j \neq i}^{n} \log \frac{(n-1)}{g_j} \mathrm{I}(w_{ij} > 0), \tag{1}$$

where W is the set of weights associated to each edge; g_j is the degree of node j, which is a neighbour of node i; $\mathrm{I}(\cdot)$ is an indicator function that equals 1 if there is an edge that connects nodes i and j with positive weight, w_{ij}.

We postulate that the degree of diversity provides relevant information about how pervasive a topic is in the weave of the economy and could ultimately attract the attention of a variety of economic actors, e.g. institutional investors, policy makers, private investors, etc..

The third dimension of the index, i.e. connectivity, assesses the weighted betweenness centrality of the ERKs, see Brandes (2001) and Freeman (1978). Connectivity measures how much a word is embedded in a discourse acting as a bridge between its parts, or more specifically, how often a word appears in-between the network paths which interconnect the other words in the text. Following Wasserman and Faust (1994), for node i we have:

$$C_i = \sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}, \tag{2}$$

where $d_{jk}(i)/d_{jk}$ is the proportion of shortest network paths connecting nodes j and k (measured by edge weights) that include the node i. Finally, an index is constructed as a composite score obtained by summing the standardized measures of prevalence, diversity and connectivity discussed above. The standardisation is carried out considering the semantic network of each time period.

3 Forecasting stock and bond markets

The objective of our paper is to assess whether, and if so to what extent, the information contained in textual news data improves the predictability of macro-finance variables. Focusing on the Norwegian stock market, Larsen and Thorsrud (2019) find strong evidence of asset prices predictability by textual news data. They use a latent Dirichlet allocation model that statistically categorises the corpus, i.e., the whole collection of words and articles, into topics that best reflect the corpus's word dependencies. Building on the aforementioned work, we rely on our more general semantic importance index, described in Section 2, as the aggregate measure of textual news data information content. Our empirical evaluation centres on the Italian stock and bond markets. This choice is not coincidental. Italy is a large economy, i.e. the 8th world largest economy by GDP, with a fairly large stock market, among the most liquid in Europe and, with the third largest sovereign debt in the world after Japan and US. This makes the Italian government debt market very attractive, and thus liquid, at all maturities. We assess the predictive power of textual news information as a non-traditional driver of the returns level, as well as volatility, of five stock and bond time series.

Our set of target variables comprises the aggregate market portfolio, i.e. FTSE MIB index, and 2, 5, 10 and 30-year maturity Italian government bond indices, collected from Datastream.

The bond indices are maintained by Refinitiv and are designed to track the performance of euro-denominated securities publicly issued by Italy for its domestic market. Such indices provide high-quality measurements of bonds with similar maturities available in the market. For each of these variables, available daily, we compute logarithmic returns and realised volatilities (measured in percentage), aggregated at a weekly frequency. We opt for a weekly aggregation frequency following evidence in Fronzetti Colladon (2020), which showed that daily news data is highly variable and proved that the effect produced by multiple news in one-week has more impact on citizens' behaviour. Indeed, investors can take time to form an opinion on newspaper contents and can adjust their allocation only after sometimes. Moreover, online news is often accessed even several days after publication and can therefore extend their influence over time.

Weekly realised volatility is computed using the range estimator of Parkinson (1980), i.e. $(4\log 2)^{-1}(\log H_t/L_t)^2$, where h_t and l_t represent the highest and lowest prices of week t, respectively. The sample period spans from January 6, 2017 to August 28, 2020, totalling 191 weekly observations.

	Returns				Realised Variance			
	Avg	SD	Max	Min	Avg	SD	Max	Min
FTSE MIB	-0.004	3.795	11.600	-36.961	0.101	0.378	0.001	5.080
BTP-2y	-0.009	0.301	-1.240	1.829	0.001	0.001	0.000	0.127
BTP-5y	0.027	0.834	-3.967	4.123	0.008	0.031	0.000	0.348
BTP-10y	0.060	1.327	-6.855	4.389	0.020	0.050	0.000	0.479
BTP-30y	0.099	2.295	-9.853	8.967	0.053	0.016	0.002	1.117

Table 1: *Descriptive statistics (mean, standard deviation, maximum and minimum) of the weekly FTSE MIB and the 2, 5, 10 and 30-year maturity BTP indices return and realised variance.*

Table 1 provides descriptive statistics for our 10 target series. Over the period analysed, stock market returns are negative on average and, as expected, more volatile than bond returns. The lowest stock market return (-37%) occurs in the week between Monday, March 9, 2020, and Friday, March 13, 2020, i.e. the inception of the COVID-19 crisis. Bond returns for maturities longer than 2 years are positive on average, likely boosted by the quantitative easing program initiated by the European Central Bank, and exhibit an upward sloped volatility term structure.

3.1 Textual data collection and key words

Choosing pertinent keywords to search in a database of newspapers articles is crucial for the construction of an informative textual index. As documented in literature, such choice is non-trivial because word meaning can vary across fields and users. For instance, Loughran and McDonald (2011) find that words associated to a negative attribute or meaning by widely used dictionaries, e.g. the Harvard Dictionary, are words typically not considered negative in financial contexts. To circumvent the problem we select a limited set of words with a clear economic meaning, homogeneously understood by the larger public.

We choose 38 sets of keywords: 24 singletons, i.e. individual words, albeit included also in plural form when it exists, and 14 sets of words sharing similar meaning, i.e. synonyms, or identifying similar items or meaning. The set of keywords is reported in Table 2, translated from Italian where needed.

Word singletons							
1	Spread	7	Quantitative easing	13	Rating	19	Real economy
2	Interest rates	8	Monetary policy	14	Eurogroup	20	European Commission
3	Euro	9	Bank of Italy	15	coronabond	21	Eurobond
4	European troika	10	ESM[1]	16	SURE[2]	22	EIB[3]
5	Junk bond	11	Oil	17	Gold	23	Financial markets
6	Strikes	12	INPS[4]	18	GDP	24	Confindustria[5]

Word sets							
1	COVID, coronavirus	5	BTP, BOT, CCT[6]	9	savings, savers	13	european union, EU
2	lockdown, quarantine	6	inflation, prices	10	deficit, gov.t debt	14	consumption, -umers
3	taxes, taxation, wealth tax	7	Borsa Italiana, FTSE MIB, FTSE MIB	11	unions, CISL, CGIL, UIL[7]		
4	economic crisis, recession, economic pandemic	8	unemployment, redundancy, unemployment benefit	12	smart working, distance work		

Table 2: [1]*European stability mechanism.* [2]*European instrument for temporary support to mitigate unemployment risks in an emergency.* [3]*European Investment Bank.* [4]*Italian social welfare and pension institution.* [5]*Italian industrial sector association.* [6]*Italian debt instruments.* [7]*Acronyms of the three largest Italian trade unions.*

The textual data used in the analysis is provided by Telpress International and it is collected from multiple online news sources. To generate networks from texts and to calculate our semantic importance index, we rely on the SBS BI web application[1], see Fronzetti Colladon and Grippa (2020), and the computing resources of the ENEA/CRESCO infrastructure (Ponti et al., 2014). Prior to the computation of the semantic importance index, common text pre-processing routines (Perkins, 2014), such as tokenisation, removal of stop-words and removal of word affixes, known as stemming (Willett, 2006), are implemented. Then, a social network based on word co-occurrences is generated for each time interval considered in the analysis.

The database of Italian news, published between January 2, 2017 and August 30, 2020, contains more than 772,500 news articles. For each news we only consider the title and the lead, i.e. the initial 30% of text, ignoring the remaining part.[2] This is consistent with previous work, which suggested that semantic importance indices are more informative when calculated on the news parts that better capture the readers' attention, i.e. the title and the lead, see Fronzetti Colladon (2020). This is also aligned with past research, which has already proven that a large part of internet users only read the beginning of online articles, see Nielsen and Loranger (2006), among others.

We calculate an index, as detailed in section 2, for each of the ERKs listed in Table 2, thus

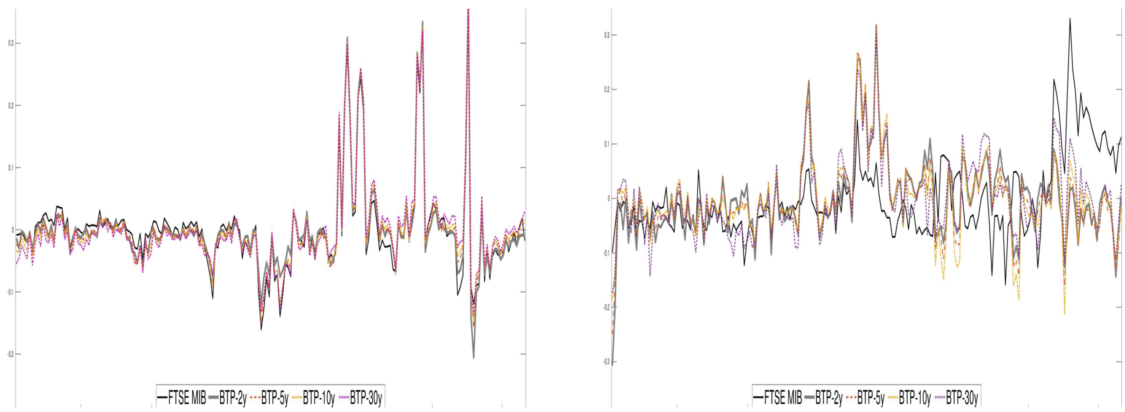
[1]Available at https://bi.semanticbrandscore.com

[2]We have also tried using the full content of news, but accuracy was not superior, computational time was substantially longer and results are not reported.

obtaining 38 time series. To reduce uncertainty and aggregate information, following Fronzetti Colladon et al. (2019), we apply Partial Least Squares (PLS) between the target variable and the (38 ERKs) predictors, incorporating information from both the definition of scores and loadings. Therefore, our measure is a series specific index.

Figure 2 shows our full sample indices associated to returns and volatilities target variables. For the sake of comparability, each index is centred and standardised. Panel A indicates that the indices targeted on the return series follow similar patterns, somewhat less evident when compared to those targeted to volatility, with large movements associated to destabilising political and economic events. For example, the sensible negative movement in the index during spring 2018 coincides with elections where no party achieved a sufficient majority; the sharp positive bounce during autumn 2018, instead, reflects the reaction to pension and social welfare reforms, introduced by the government composed by the conservative LEGA and populist M5S parties, that markets and the European Union did not fully support. Similar patterns are observed during the late spring and though summer 2019, period in which the Italian government suffered strong internal disagreements on several affairs, including immigration, and which resolved in a new government coalition. Finally, the effect of the inception of the COVID-19 pandemic is clear.

Panel B shows that indices targeted to volatility are more variable than those constructed for returns. Indeed, the average correlation of the latter is 97%, the average correlation of volatility targeted indices is 47%. However, the occurrence of large shocks and instability is aligned between the two sets of indices. The only striking exception is observed at the end of 2019, when patterns of the two sets of indices series appears to diverge.



(a) Media news index targeted to the returns series *(b) Media news index targeted to the volatility series*

Figure 2: *Panel A shows the semantic importance indices applied to Italian stock returns (FTSE MIB) and Italian bond returns (BTP-2y, BTP-5y, BTP-10y, BTP-30y). Panel B shows the indices applied to volatility of Italian stock markets and Italian bond markets.*

3.2 Design of the forecasting exercise

Our forecasting exercise aims at assessing whether including information stemming from news contributes to improving stocks and bonds returns predictions. We employ a recursive forecasting scheme, using an expanding estimation sample, to produce 1-week ahead forecasts. The first estimation sample spans from 6 January 2017 to 26 April 2019. The out-of-sample (OOS) forecast evaluation period spans the following 70 weeks, i.e. from 3 May 2019 to 28 August 2020. Therefore, we consider both the pre-COVID19 period, the turbulent COVID-19 outbreak period in March and April 2020, and the following less volatility COVID-19 period up to the end of summer 2020.

We opt for a simple forecasting model, i.e. the ARX(1):

$$y_{t+1} = \alpha + \gamma y_t + \beta x_t + \varepsilon_{t+1} \tag{3}$$

where y_{t+1} is the target variable we aim at predicting, x_t is a set of news information predictors and $\varepsilon_{t+1} \sim WN(0, \sigma^2)$. An obvious choice for x_t is the pool (or a subset) of the 38 index assigned to the ERKs listed in Table 2. If, on the one hand, this approach enables identifying and isolating ERKs with predictive power from the rest, on the other hand, the large set of regressors, relative to the limited number of observations in our sample, may generate an undesirable level of uncertainty.

To circumvent this problem we convey the information contained in the index associated with the individual (sets of) keywords using two alternative aggregation approaches. The first approach extracts, from the pool of 38 ERKs variables, one or more common factors by means of partial least square (PLS).[3] The PLS is computed individually for each series and repeated for each vintage that forecasts are produced, therefore using real-time information. We labelled it as "ERK model" in the remainder of the paper. The competing aggregation method, following Stock and Watson (1999) and Timmermann (2006), consists in computing multiple forecasts for each target variable using Equation (3) and one ERK predictor, $x_{i,t}$, at the time and then in aggregating those forecasts. We use the equal weight combination:

$$y_{t+1} = \sum_{i=1}^{38} w_i \hat{y}_{i,t+1} \tag{4}$$

where $\hat{y}_{i,t+1}$ is the forecast of y_{t+1} generated by the linear ARX(1) using $x_{i,t}$ and $w_i = 1/38$, $i = 1, \ldots, 38$.

We contrast the predictive performance of the model exploiting textual news information against two standard benchmarks: the white noise model when the target is the return of either stocks and bonds and the AR(1) model when we aim at forecasting volatility.[4] Welch and Goyal (2008) document how difficult it is to provide a superior forecast performance than the white noise model when predicting stock returns. De Pooter et al. (2010) show similar evidence for bond markets. This is because prices or yields are non-stationary, or close to, which makes it difficult to outperform the simple no-change forecast. When we turn to volatility instead, the non-stationary component is typically less relevant and the time-reversion is more pronounced, with periods of high and low volatility alternating.[5]

Moreover, we also compare our index to an alternative and well-known text evaluation method:

[3] We have also tried principal component analysis (PCA), but results were inferior and not reported. We believe that the total number of ERKs is limited and PCA is less adequate in this case.

[4] The white noise model for returns corresponds to the driftless random walk model for price. We refer to it, labelling it RW, in the remaining part of the paper.

[5] We have also applied the RW no-change benchmark to volatility predictions, results are substantially inferior to the AR models and not reported in the text.

the sentiment index, see Fraiberger et al. (2018). This index is computed by evaluating media sentiment in correspondence with economy-related terms, considering the online news articles included in our study. Prior to the calculation of sentiment, we apply the same text pre-processing procedures described in 3.1. Subsequently, we consider the polarity of the words associated to each ERK, and compute a weighted average based on co-occurrence strength.[6] Accordingly, the sentiment index may vary from -1 to +1, where negative values represent negative expressions and positive values represent positive expressions. We substitute our textual indicator with the sentiment index in Equation (3) and apply it as an alternative news information predictor. We label it "SI model" in the remainder of the paper.

We measure forecast accuracy by means of mean squared prediction errors (MSPEs). The statistical assessment of the predictive performance differential stemming from the inclusion of textual news information is based on the Diebold and Mariano test (Diebold and Mariano, 1995, DM).

3.3 Results

In-sample evidence Results from Inoue and Kilian (2004) imply that in-sample predictability is a necessary condition for out-of-sample predictability. To assess the degree of in-sample fit, in Figure 3, we compare the ERK model based on the semantic importance index to the nested benchmark model by means of Akaike information criterion (AIC). See, for example, Ravazzolo and Rothman (2013) for a similar exercise on the role of oil prices for predictability to US output growth.[7] The AIC is computed recursively for all estimation windows between 3 May 2019 and 28 August 2020.

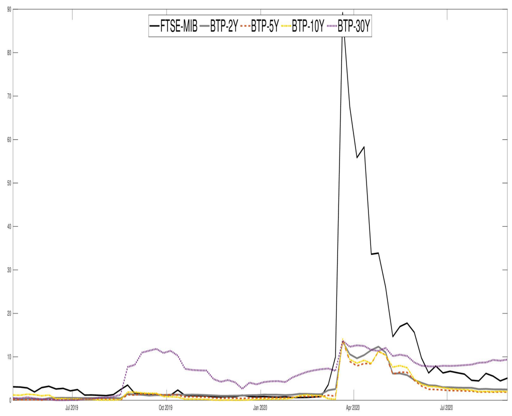
The inclusion of textual news information improves stock and bond markets predictability in all the sample. The gains are moderate in the first part of the sample. Starting from March 2020, i.e. the inception of the COVID-19 crisis, until May 2020, for all the series, the ERK model exhibits a dramatic drop in the AIC, resulting in large positive values for the difference that indicate a large predictability power from our index. The largest difference is observed when predicting stock market returns. Gains continue to exist in the second part of the COVID-19 period after May 2020 associated to lower volatility in the series. The ERK model shows the most persistent gains over the sample for the bond with the longest maturity.

The data shows a similar picture when turning to stock market and bond volatility, albeit less consistent. Until March 2020, for the stock market returns, the ERK is inferior or similar to the benchmark, but then it dominates for the remaining part of the sample. When turning to bond volatility, in general, the ERK model performs well from summer 2019, a period characterised by political uncertainty in Italy. The ERK performance stabilises during autumn 2019 after a new

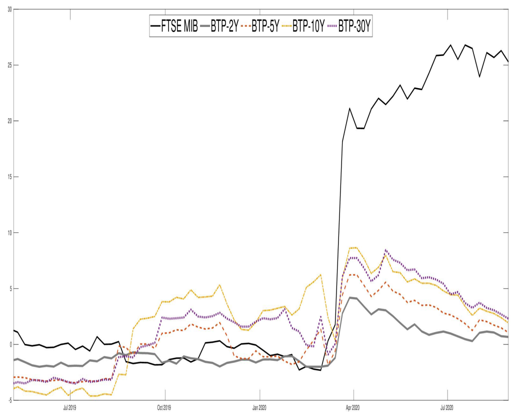
[6] We additionally tested other approaches for the calculation of sentiment, such as considering the full articles' content and not just the sentences related to economic terms. However, none of these alternative approaches led to results better than our primary choice.

[7] Note that the benchmark model when predicting volatility series is the AR(1). For returns, the benchmark is the white noise process. However, because the latter does not have any explanatory component, its AIC is very poor, i.e. it is a function only of the series variance, see the forecasting puzzle in Meese and Rogoff (1983). Therefore, for the sake of a fair comparison, we contrast the AR(1) and ERK models in all ten evaluations.

government was formed and markets experienced a period of calm until the COVID-19 crisis. The gains are lower in the second part of the COVID-19 period, but the difference remains positive for all four bond comparisons.



(a) Returns series



(b) Volatility series

Figure 3: *Panel A shows differences in AIC (AIC(AR) - AIC(ERK)) for the AR model without the semantic importance index and the alternative ARX model with the semantic importance index (ERK) when predicting Italian stock returns (FTSE MIB) and Italian bond returns (BTP-2y, BTP-5y, BTP-10y, BTP-30y); if the AR model generates the better fit, then the AIC differences are negative. Panel B shows differences in AIC (AIC(AR) - AIC(ERK)) for the AR model and the alternative ERK when predicting volatility of Italian stock and bond markets.*

Returns out-of-sample results Results in Panel A of Table 3 show that the ERK model that includes the semantic importance of economic related keywords improves forecasting accuracy for the shorter and longer maturity bond returns. In particular, reduction in MSPE for the 2-year bonds is statistically significant at 10%. MSPE reduction for the 30-year bonds is more moderate and the performance of the RW and ERK models are more similar for the 5-year and 10-year maturity. In the case of stock returns (FTSE MIB), results are comparable to the benchmark, but not superior. Studying performance over time, the ERK model gives the largest predictability during the first part of the COVID-19 period, but this predictability smooths in the second part of the COVID-19 period for the stock returns and bond returns with middle term maturities, confirming AIC evidence in Figure 3.

The AR model never provides comparable results and only for the 2-year BTP maturity does it achieve lower MSPE than the RW. The forecast combination of ARX (EW) is also statistically superior to the benchmark for the 2-year bond returns, yet falls behind the ERK model. Finally, in none of the five cases does the model based on the sentiment index offer gains similar to the ERK model and it is never superior to the benchmark RW. Therefore, how newspaper information is treated and modelled matters.

Volatility out-of-sample results Panel B of Table 3 provides results on volatility predictions. In this case, we observe the largest gains, in terms of forecast accuracy, obtained by using the semantic importance index. In all the five cases, the ERK model is statistically superior to the

AR benchmark, with MSPE reductions ranging from 9% to 1%. The large improvement is when forecasting stock market volatility, with a smaller MSPE of 9% and statistically significant at 1% confidence level. When predicting bond volatility, the ERK statistically outperforms the benchmark for all maturities, with the largest economic gains observed for the 10-year maturity (6% improvement). The forecast combination does not measure up, often resulting statistically inferior to the benchmark. In addition, the SI model does not perform similarly to the ERK: SI gives lower MSPE than the AR for the volatility of the stock index and the 30-year bond returns, but the reduction is never statistically significant. This result stresses the importance of the index construction to avoid an unfavourable signal to noise balance. Indeed, textual data helps to improve forecast accuracy provided that specific keywords that receive more attention in newspapers are selected.

Models	FTSE MIB		BTP-2y		BTP-5y		BTP-10y		BTP-30y	
Panel A: Returns										
RW	3.07		0.18		0.59		1.02		1.89	
AR	1.07		0.99		1.00		1.01		1.00	
ERK	1.02		0.98	*	1.02		1.01		0.99	
EW	1.06		0.98	*	1.03		1.05		1.05	
SI	1.07		1.00		1.01		1.01		1.00	
Panel B: Volatility										
AR	0.92		0.94		0.88		0.85		0.80	
ERK	0.91	**	0.97	*	0.96	*	0.94	**	0.99	*
EW	1.00		1.00		1.02		1.02		1.01	
SI	0.95		1.00		1.00		1.00		0.99	

Table 3: *Panel A provides mean square prediction error (MSPE) results when forecasting Italian stock returns (FTSE MIB) and Italian bond returns (BTP-2y, BTP-5y, BTP-10y, BTP-30y). Absolute MSPE for the random walk (RW) benchmark is reported; relative numbers to the benchmark are given for the alternative autoregressive (AR), autoregressive extended with the semantic importance index (ERK), equal weight combination of ARX models based on different economic related keywords (EW) and autoregressive extended with the sentiment index (SI). Panel B provides MSPEs when forecasting the volatility of Italian stock and bond markets. Absolute MSPE for the AR benchmark is reported; relative numbers to the benchmark are given for the alternative specifications. One $*$ and two $**$ indicate that the alternative model provides superior statistical forecasts at* 10 *and* 5 % *significance level, respectively.*

4 Conclusion

This paper introduces a new textual data index for predicting stock market data. The index is based on a novel methodology applied to a large set of newspaper articles to evaluate the importance of one or more general economic related keywords that appear in a text. The index considers three dimensions: prevalence, connectivity and diversity. The methodology is applied to online Italian press and 38 economic related keywords are selected. The resulting index is used to predict the Italian stock market and government bond returns and volatilities in the 2017-2020 period, including the inception of the COVID-19 crisis.

Our findings show that the semantic importance index based on media news text data is able to capture the different phases and individual features of return and volatility dynamics of financial variables. Periods of large movements in the index are associated with political and economic instability. When used to predict weekly market and bond returns and volatilities, we find strong evidence of predictability of bond returns and volatility, as well as of stock market volatility.

References

Baker, M. and Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, 61(4):1645–1680.

Baker, M. and Wurgler, J. (2007). Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, 21(2):129–152.

Beaudry, P. and Portier, F. (2006). Stock Prices, News, and Economic Fluctuations. *American Economic Review*, 96(4):1293–1307.

Beaudry, P. and Portier, F. (2014). News-Driven Business Cycles: Insights and Challenges. *Journal of Economic Literature*, 52(4):993–1074.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *The Journal of Mathematical Sociology*, 25(2):163–177.

Chung, S.-L., Hung, C.-H., and Yeh, C.-Y. (2012). When Does Investor Sentiment Predict Stock Returns? *Journal of Empirical Finance*, 19(2):17 – 240.

de Jong, S. (1993). SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251 – 263.

De Pooter, M., Ravazzolo, F., and van Dijk, D. (2010). Term Structure Forecasting using Macro Factors and Forecast Combination. Technical report, International Finance Discussion Papers Number 993.

Diebold, F. X. and Mariano, R. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13:253–265.

Fraiberger, S. P., Lee, D., Puy, D., and Rancière, R. (2018). Media sentiment and international asset prices. Working Paper 25353, National Bureau of Economic Research.

Freeman, L. C. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks*, 1(3):215 – 239.

Fronzetti Colladon, A. (2018). The Semantic Brand Score. *Journal of Business Research*, 88:150 – 160.

Fronzetti Colladon, A. (2020). Forecasting Election Results by Studying Brand Importance in Online News. *International Journal of Forecasting*, 36(2):414 – 427.

Fronzetti Colladon, A. and Grippa, F. (2020). Brand intelligence analytics. In Przegalinska, A., Grippa, F., and Gloor, P. A., editors, *Switzerland*, volume 1 of *Digital Transformation of Collaboration*, pages 125–141. Springer Nature.

Fronzetti Colladon, A., Grippa, F., and Innarella, R. (2020). Studying the association of online brand importance with museum visitors: An application of the semantic brand score. *Tourism Management Perspectives*, 33:580 – 588.

Fronzetti Colladon, A., Guardabascio, B., and Innarella, R. (2019). Using Social Network and Semantic Analysis to Analyze Online Travel Forums and Forecast Tourism Demand. *Decision Support Systems*, 123:1–11.

Fronzetti Colladon, A. and Naldi, M. (2020). Distinctiveness Centrality in Social Networks. *PLoS ONE*, 15(5):163–177.

Inoue, A. and Kilian, L. (2004). In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econometric Reviews*, 23:371 – 402.

Jussila, J., Vuori, V., Okkonen, J., and Helander, N. (2017). Reliability and Perceived Value of Sentiment Analysis for Twitter Data. In Kavoura, A., Sakas, D., and Tomaras, P., editors, *Strategic Innovative Marketing*, Springer Proceedings in Business and Economics. Springer.

Keller, K. L. (1993). Conceptualizing, Measuring, and Managing Customer-Based Brand Equity. *Journal of Marketing*, 57(1):1 – 22.

Larsen, V. and Thorsrud, L. (2019). The Value of News for Economic Developments. *Journal of Econometrics*, 210(1):203 – 218.

Limongi Concetto, C. and Ravazzolo, F. (2019). Optimism in Financial Markets: Stock Market Returns and Investor Sentiments. *Journal of Risk Financial Management*, 12(85):1–14.

Loughran, T. and McDonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35 – 65.

Meese, R. and Rogoff, K. (1983). Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample? *Journal of International Economics*, 14(1-2):3 – 24.

Nielsen, J. and Loranger, H. (2006). *Prioritizing Web Usability*. Pearson Education, Birmingham, UK.

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *Journal of Business*, 53:61–65.

Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing, Birmingham, UK.

Piantadosi, S. T. (2014). Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Ponti, G., Palombi, F., Abate, D., Ambrosino, F., Aprea, G., Bastianelli, T., Beone, F., Bertini, R., Bracco, G., Caporicci, M., Calosso, B., Chinnici, M., Colavincenzo, A., Cucurullo, A., D'angelo, P., Rosa, M. D., De Michele, P., Funel, A., Furini, G., and Vita, A. (2014). The role of medium size facilities in the HPC ecosystem: the case of the new CRESCO4 cluster integrated in the ENEAGRID infrastructure.

Ravazzolo, F. and Rothman, P. (2013). Oil and U.S. GDP: A Real-Time Out-of-Sample Examination. *Journal of Money, Credit and Banking*, 45(2-3):449–463.

Stock, J. H. and Watson, W. M. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44:293–335.

Timmermann, A. (2006). Forecast Combinations. In Elliot, G., Granger, C., and Timmermann, A., editors, *North-Holland*, volume 1 of *Handbook of Economic Forecasting*, chapter 4, pages 135–196. Elsevier.

Wasserman, S. and Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press, New York, NY.

Welch, I. and Goyal, A. (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Willett, P. (2006). The Porter Stemming Algorithm: Then and Now. *Program Electronic Library and information systems*, 40:219–223.

CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

The Center for Research in Economics and Statistics (CREST) is a leading French scientific institution for advanced research on quantitative methods applied to the social sciences.

CREST is a joint interdisciplinary unit of research and faculty members of CNRS, ENSAE Paris, ENSAI and the Economics Department of Ecole Polytechnique. Its activities are located physically in the ENSAE Paris building on the Palaiseau campus of Institut Polytechnique de Paris and secondarily on the Ker-Lann campus of ENSAI Rennes.