LINKING ASSET PRICES TO NEWS WITHOUT DIRECT ASSET MENTIONS

[1]ILANIT AVIOZ, [2]HAIM KEDAR-LEVY, [3]CRINA PUNGULESCU, [4]DAVID STOLIN

[1,2]School of Management, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva 94105, Israel
[3]Institute for Economic Forecasting, National Institute of Economic Research and Romanian Academy, Casa Academiei, Calea 13 Septembrie Nr. 13, 050711 Bucharest, Romania
[4]TBS Business School, 1 Place Alphonse Jourdain - CS 66810 31068 Toulouse Cedex 7, France
E-mail: [1]aviozi@post.bgu.ac.il, [2]hlevy@bgu.ac.il, [3]crinapungulescu@ipe.ro, [4]d.stolin@tbs-education.fr

Abstract - Advances in Natural Language Processing (NLP), computing power and data availability are driving an explosion in research about the impact of news on asset prices. However, when relating news to individual assets, this research is based on mentions of specific assets or related terms in the news stories. Such an approach has two shortcomings. First, it requires a substantial time investment in a specific NLP technology. Second, and more importantly, it ignores news articles that do not directly mention a given asset or a pre-defined asset-related term, even if these articles are logically related to the asset in question. Our approach relies instead on a novel NLP technology called "semantic fingerprinting", which projects any text onto a binary vector representing its meaning. The greater the overlap between the semantic fingerprint of a news article and a given asset description, the more relevant we expect the article to be, whether or not the given asset is mentioned in the news directly. We show that this approach successfully picks up the positive impact of news on prices of commonly traded commodities using a dataset of general news published by The Guardian. We include the needed data and instructions for implementing this approach.

JEL classification: G11, G15

Keywords - News, Volatility, Commodities, Natural Language Processing

I. INTRODUCTION

Existence of the link between news coverage and asset prices is well established (Tetlock, 2011), and is of first-order importance to academics as well as practitioners. However, research has tended to focus on news that directly mention the assets in question. While news stories can and do impact assets which are not mentioned in them, studying these links typically requires from a researcher a very significant time investment into learning and implementing an advanced Natural Language Processing (NLP) technology. In this paper, we show how a very easy to understand and to implement NLP technology allows us to link asset price movements to related news that do not mention the asset directly. The technology in question is semantic fingerprinting (De Sousa Webber, 2015). After constructing a fixed semantic map based on a large corpus of Wikipedia articles, it can then project any word or text onto that map, with 1's indicating the location of related terms, and 0's, unrelated terms. Thus, overlap between the semantic fingerprints of a news item and an asset's name is indicative of their semantic connectedness. We exploit this insight to show how news impact the price volatility of major commodities – even after news stories mentioning a given commodity by name are excluded. This approach significantly lowers the barrier to the study of how news impact prices of assets that are not directly mentioned in them. To this end, we include instructions for other researchers interested in implementing this approach The rest of the paper is organized as follows. The next section presents our methodology while Section 3 overviews our data. Section 4 presents our results and Section 5 concludes.

II. METHODOLOGY

Our working hypothesis is straightforward: we expect that related news coverage has an impact on volatility. In particular, it is natural to expect that the more news coverage there is, the greater would the volatility be (see Bystr¨om, 2016, for instance).

In order to operationalize our tests of the impact of news on returns, we require the following elements:

1. An efficient approach to processing a large body of news text. We do so by using the semantic fingerprinting methodology.
2. Once the news text has been processed, we need a measure to capture the volume of relevant news. We call this measure news intensity.
3. Once news intensities have been obtained, we need an econometric approach to study their impact on returns.

Semantic fingerprinting, news intensities and our econometric approach are described below.

2.1 Semantic Fingerprinting

Semantic fingerprinting as implemented by Cortical.io (De Sousa Webber, 2015) represents text as a sparse binary 128 × 128 matrix of 16,384 topics or "contexts". The semantic fingerprint of a word is

represented by the "contexts" with which the word is associated in a training corpus. The semantic fingerprint of a text is the aggregation of semantic fingerprints of key words in that text. Ibriyamova et al. (2017, 2019) demonstrate semantic fingerprinting to be an effective document comparison tool in the investment context. In particular, they show that similarity of company descriptions, measured as the cosine similarity of their semantic fingerprints, predicts the similarity of their stock returns. In this project, we take semantic fingerprints of the 29 commodities in our sample, and use them as filters for the semantic fingerprints of news stories, in order to focus on news most relevant to each commodity. In addition to picking up explicit references in news stories, this approach attributes to a commodity news stories that are likely to be of relevance even if the commodity itself is not mentioned in them; for example, the semantic fingerprints of energy commodities will have greater overlap with the semantic fingerprints of news related to the countries producing them. In this way, we operationalize the volume of relevant news as the news intensity for any commodity/day combination.

2.2 News Intensities

More formally, let N(t) be the total number of news stories on day t. Corresponding to each news story is a semantic fingerprint comprising a subset of K possible positions of the semantic space, each of them representing a group of related terms. Let $J_{n,t}$ be a K−vector containing the semantic fingerprint of the nth news story on day t. Further, we define I_i as a K−vector containing the semantic fingerprint of the ith commodity. Then, the news intensity score for commodity i on day t is defined as the average of the number of news stories whose fingerprints include a given position, taken across all positions comprising commodity i's fingerprint, and scaled by the average number of stories across all possible positions:

$$NI_{i,t} = \frac{\left(\Sigma_{n=1}^{N(t)} I_i \cdot J_{n,t}\right) / ||I_i||}{\Sigma_{n=1}^{N(t)} ||J_{n,t}||/K} \quad (1)$$

where (.) is the dot product operator, and (|| ||) is the vector length. As a simple illustration, suppose that we use a semantic space comprised of only K = 2 × 2 = 4'

positions. Suppose further that commodity A's fingerprint is (1,1,0,0), and commodity B's is (0,1,1,1), and that on day t there are N(t) = 5 news stories whose fingerprints are (1,0,0,0), (1,1,0,0), (1,1,1,0) and (1,0,1,0) and (0,1,0,1). Then, the aggregate semantic fingerprint across the five news stories is (4,3,2,1) and the total number of

semantic positions across the five news stories is $\Sigma_{n=1}^{N(t)} J_{n,t}$ = 4+3+2+1 = 10, giving a measure of the "newsworthiness" of day t. Of them, $\Sigma_{n=1}^{N(t)} I_A \cdot J_{n,t}$ = 4 + 3 = 7 positions overlap with A's fingerprint, and $\Sigma_{n=1}^{N(t)} I_B \cdot J_{n,t}$ = 3+2+1 = 6 positions overlap with B's fingerprint. Therefore, the number of relevant news fingerprint positions per Commodity

A fingerprint position is $\left(\Sigma_{n=1}^{N(t)} I_A \cdot J_{n,t}\right) / ||I_A|| = 7/2 = 3.5$; the analogous quantity for Commodity B is $\left(\Sigma_{n=1}^{N(t)} I_B \cdot J_{n,t}\right) / ||I_B|| = 6/3 = 2$. Dividing these by the scaling factor of $\Sigma_{n=1}^{N(t)} ||J_{n,t}||/K = 10/4 = 2.5$ produces news intensities of $NI_{A,t}$ = 3.5/2.5 = 1.4 and $NI_{B,t}$ = 2/2.5 = 0.8. Their interpretation is that the amount of news per fingerprint position for Commodity A is 40% higher and for Commodity B, 20% lower than the average amount of news for a randomly chosen fingerprint position.

2.3 Econometric Approach

Following the GARCH framework described in Hansen and Lunde (2005), we define the continuously compounded daily return as $r_{t+1} = \ln(p_{t+1}) - \ln(p_t)$, where p_t are daily commodity prices and to model the conditional density of returns, we choose a constant mean model for the mean equation, the exponential GARCH of Nelson (1991) for the variance equation and the t−distribution for the density function of the standardized residuals.

Specifically, to model the conditional variance, σ_{t+1}^2, we augment the exponential GARCH(1,1) with our measure of news intensity, NI, as an external regressor, as follows:

$$\ln\left(\sigma_{t+1}^2\right) = \omega + \zeta NI_t + \alpha e_t + \gamma\left(|e_t| - E|e_t|\right) + \beta \ln\left(\sigma_t^2\right) \quad (2)$$

The focus of our paper are the $\hat{\zeta}$ coefficients which convey the impact of the newsintensity on asset returns volatility.

III. DATA

We obtain price data for all the 29 commodities indices available from Bloomberg: aluminium, Brent crude oil, cocoa, coffee, copper, corn, cotton, feeder cattle, gas oil, gold, heat oil, lead, lean hogs, live cattle, natural gas, nickel, orange juice, platinum, silver, soybean meal, soybean oil, soybeans, sugar, tin, unleaded gas, wheat, Kansas wheat, WTI crude oil and zinc. Our news dataset consists of 44,817 articles published in The Guardian daily in the period 1.11.2019 − 1.05.2020.

At this point, a necessarily subjective decision has to be made about the asset descriptions to be fingerprinted by Cortical.io's AI solution, the Retina engine. De Sousa Webber (2015) notes that fingerprints of shorter texts favor synonyms as opposed to fingerprints of longer texts, which uncover mainly associated concepts. Ibriyamova et al. (2019) show that short text descriptions of stocks perform better than long 10-K S.E.C. filings for predicting stock return correlations. Given that most commodities can be unambiguously described in one word (e.g. "gold", "cotton", etc.) we make use of the facility of the Retina engine to provide short fingerprints of 328 "contexts" for any given term it knows. That means that we choose to fingerprint a one-word term for the commodities that are described typically by two words: "gas" for natural gas and unleaded gas, "oil" for Brent crude oil, WTI crude oil, heat oil and gas oil, "oranges" for orange juice, "soybeans" for soybean meal, soybean oil and soybean, "cattle" for feeder cattle and live cattle and "hogs" for lean hogs.

Using the commodities fingerprints in conjunction with the fingerprints of each of the daily articles published in The Guardian over the period, we calculate the measures of news intensities corresponding to each of the 29 commodities indices as described above. Table 1 presents descriptive statistics for our return data as well as the corresponding series of news intensities, which appear substantially more present in the news than a randomly chosen fingerprint position.

IV. EMPIRICAL RESULTS

Table 2 (Panel A) shows how our measure of news intensity contributes as an external regressor in the variance equation of the exponential GARCH(1,1) models estimated for our the 29 commodity indices. The results are overwhelmingly in favor of a positive impact of news on volatility: the coefficients are positive for 28 commodities out of 29 and statistically significant at 5% for 19 of them. To illustrate the fact that semantic fingerprinting precludes the need of pre-defining lists of asset relevant terms, we reestimate our models on a reduced news dataset for each commodity by eliminating the articles that mention it directly. Panel B of Table 2 confirms that our results are not driven by the news items that mention directly our assets of interest, but rather by the news related to the "contexts" that constitute the semantic fingerprint of each commodity. Partial news effects are computed by multiplying the average daily change in news intensities (in absolute values), $|\Delta NI_t|$, with the estimated $\hat{\zeta}$ coefficients and reported, to facilitate interpretation, as percentages of the logarithm of the variance of returns in our sample in Table 2 . The news impact for the 29 commodities in our sample is on average 1.21%-1.22% relative to the logarithm of the variance of asset returns.

For a further illustration of the advantages of semantic fingerprinting over word lists we obtain a description of the nickel commodity market, extract keywords using Cortical.io's solutionand count the number of articles in our news database that mention these relevant terms. The keywords are: "nickel" (14 articles), "ore(s)" (123 articles), "prices" (1589 articles), "metal(s)" (695 articles), "steel" (649 articles), "demand" (5741 articles), "alloy(s)" (41 articles), "oxide" (31 articles) and "mining" (1405 articles). While this approach is successful and the useful articles outweigh the few false positives (i.e. nickel commemorative coins), it still uses less than a quarter of our news database. By contrast, semantic fingerprinting casts a wider net by design and is therefore able to extract more indirect yet potentially relevant information.

V. CONCLUSION

Relying on an easy-to-use natural language processing technology called semantic fingerprinting, we show that price volatility of commodities is strongly related to the intensity of their connection to concurrent news. What is more, this result stays even if we exclude any article that mentions a given commodity by name. Because of the efficiency of semantic fingerprint technology, this is achieved quickly for a very large news archive. In subsequent work, we expect to apply this technology to other types of assets, as well as to news across languages other than English.

REFERENCES

[1] Bystro¨m, H. (2016). Language, news and volatility. Journal of International Financial Markets, Institutions and Money 42, 139–154.

[2] De Sousa Webber, F. (2015). Semantic folding theory and its applications in semantic fingerprinting. Cortical.io White Paper, available from

[3] https://www.cortical.io/static/downloads/semantic-folding-theory-white-paper.pdf .

[4] Hansen, P. and A. Lunde (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? Journal of Applied Econometrics 20, 873–889.

[5] Ibriyamova, F., S. Kogan, G. Salganik-Shoshan, and D. Stolin (2017). Using semantic fingerprinting in finance. Applied Economics 49, 2719–2735.

[6] Ibriyamova, F., S. Kogan, G. Salganik-Shoshan, and D. Stolin (2019). Predicting stock return correlations with brief company descriptions. Applied Economics 51, 88–102.

[7] Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59, 347–370.

[8] Pungulescu, C. (2021). Bilateral home bias: A new measure of proximity. Working Paper, available from https://papers.ssrn.com/sol3/papers.cfm?abstract id=2697490.

[9] Tetlock, P. (2011). All the news that's fit to reprint: Do investors react to stale information? The Review of Financial Studies 24, 1481–1512.

This table reports descriptive statistics for the news intensities scores (mean and standard deviation) as well as for the continuously compounded daily returns of the commodities indices (mean, mean of absolute value returns, standard deviation and logarithm of variance).

	News Intensities		Returns			
Commodity	**Mean**	**StdDev**	**Mean**	**MeanAbs**	**StdDev**	**Ln(Variance)**
Aluminium	2.69	0.11	-0.000306	0.008617	0.0114	-8.9464
Brent Crude	3.71	0.24	-0.000214	0.016342	0.0247	-7.4040
Cocoa	3.52	0.18	0.000233	0.013652	0.0174	-8.1053
Coffee	4.95	0.20	-0.000093	0.014447	0.0190	-7.9236
Copper	3.94	0.13	-0.000441	0.009787	0.0131	-8.6758
Corn	4.01	0.18	-0.000019	0.010261	0.0142	-8.5131
Cotton	3.93	0.18	-0.000294	0.010236	0.0138	-8.5632
Feeder Cattle	4.39	0.20	0.000112	0.009377	0.0128	-8.7216
Gas Oil	3.71	0.24	-0.000089	0.014856	0.0214	-7.6920
Gold	3.20	0.06	-0.000298	0.006435	0.0093	-9.3528
Heat Oil	3.71	0.24	-0.000087	0.015016	0.0215	-7.6765
Lead	4.02	0.13	-0.000189	0.011469	0.0149	-8.4126
Lean Hogs	2.97	0.07	0.000320	0.016154	0.0212	-7.7049
Live Cattle	4.39	0.20	0.000092	0.008614	0.0118	-8.8824
Natural Gas	3.75	0.18	0.000658	0.020155	0.0278	-7.1655
Nickel	3.07	0.14	-0.000505	0.013482	0.0177	-8.0653
Orange Juice	3.60	0.11	0.000254	0.014267	0.0185	-7.9765
Platinum	3.24	0.13	-0.000037	0.012354	0.0172	-8.1231
Silver	3.35	0.06	-0.000248	0.011785	0.0179	-8.0445
Soybean Meal	2.47	0.22	0.000009	0.009789	0.0133	-8.6454
Soybean Oil	2.47	0.22	-0.000364	0.010393	0.0137	-8.5747
Soybeans	2.47	0.22	-0.000132	0.008547	0.0118	-8.8761
Sugar	3.72	0.23	0.000056	0.013388	0.0174	-8.1017
Tin	3.81	0.13	-0.000808	0.008750	0.0127	-8.7372
Unleaded Gas	3.75	0.18	-0.000044	0.016606	0.0261	-7.2891
Wheat	3.69	0.21	0.000011	0.012636	0.0162	-8.2437
Kansas Wheat	3.69	0.21	0.000115	0.013094	0.0169	-8.1638
WTI Crude Oil	3.71	0.24	0.000223	0.018170	0.0293	-7.0577
Zinc	2.80	0.19	-0.000516	0.011630	0.0151	-8.3891

Table 1: Descriptive Statistics

This table reports the estimated coefficients, p-values and partial effects of the news intensity external regressor, NI, in the variance equation of the exponential GARCH(1,1) model: $\ln\left(\sigma_{t+1}^2\right) = \omega + \zeta NI_t + \alpha e_t + \gamma\left(|e_t| - E|e_t|\right) + \beta \ln\left(\sigma_t^2\right)$. The coefficients that are significant at 10% appear in bold font. Panel A uses news intensities scores computed using the entire news dataset at our disposal. Panel B uses news intensities scores computed for each commodity after eliminating all the articles that mention directly that particular commodity.

The partial news effects are calculated as $\hat{\zeta} \times |\Delta NI|$, where $|\Delta NI|$ is the average of the daily changes in the news intensities scores (in absolute values) and reported as percentages of the logarithm of the variance of returns (in absolute value) for each commodity index.

	A. Full News Dataset			B. Reduced News Dataset		
Commodity	$\hat{\zeta}$	**p − value**	**Partial News Effect (%)**	$\hat{\zeta}$	**p − value**	**Partial News Effect (%)**
Aluminium	-0.23	0.39	-0.26	-0.26	0.38	-0.26
Brent Crude	1.79	0.00	3.77	1.89	0.00	3.61
Cocoa	0.14	0.00	0.20	0.13	0.00	0.17
Coffee	0.02	0.74	0.03	0.01	0.83	0.02
Copper	0.60	0.00	0.65	0.60	0.00	0.65

Corn	0.82	0.16	1.28	1.00	0.11	1.54
Cotton	1.67	0.00	2.52	1.69	0.00	2.54
Feeder Cattle	0.12	0.75	0.15	0.11	0.77	0.13
Gas Oil	1.26	0.00	2.55	0.85	0.00	1.56
Gold	2.90	0.00	1.89	2.89	0.00	1.86
Heat Oil	0.81	0.00	1.66	0.83	0.00	1.53
Lead	0.89	0.94	1.14	0.85	0.14	1.00
Lean Hogs	2.88	0.00	2.00	2.75	0.00	1.90
Live Cattle	1.17	0.02	1.46	1.01	0.08	1.25
Natural Gas	0.25	0.27	2.73	0.44	0.27	0.42
Nickel	0.09	0.00	0.10	0.16	0.00	0.18
Orange Juice	2.14	0.00	2.25	2.15	0.00	2.30
Platinum	1.05	0.00	1.55	1.12	0.00	1.63
Silver	0.67	0.83	0.48	0.48	0.74	1.03
Soybean Meal	0.20	0.00	0.29	0.25	0.00	0.36
Soybean Oil	0.08	0.00	0.11	0.10	0.00	0.14
Soybeans	0.45	0.49	0.63	0.46	0.47	0.65
Sugar	0.22	0.22	0.42	0.31	0.08	0.56
Tin	1.15	0.00	1.15	1.17	0.00	1.16
Unleaded Gas	1.64	0.15	2.88	2.28	0.00	3.40
Wheat	0.18	0.00	0.32	0.19	0.00	0.32
Kansas Wheat	0.41	0.00	0.72	0.40	0.00	0.68
WTI Crude Oil	2.26	0.00	5.00	2.35	0.40	4.72
Zinc	0.06	0.00	0.07	0.13	0.00	0.16

Table 2: Impact of News Intensity on Volatility