Università Commerciale
Luigi Bocconi

Using Sentiment Indicators in Asset Management: Big Data, Deep Learning, and Artificial Intelligence

Prof. Massimo Guidolin

20135 – Theory of Finance, Part I

(Sept. –October)

Fall 2019

Outline and objectives

- Machine learning and textual information processing

- Heston and Sinha (2017)

- Beckers (2019)

Textual information research

- Machine learning and textual information processing have become a growing part of financial practice
  - Applications concern stock picking, bankruptcy prediction, corporate distress diagnosis, consumer credit risk, to some extent option pricing
- Internet news sources and social media provide a growing universe of textual information, including internet searches, Facebook networks, and Twitter broadcasts
- Tetlock's (2007) studies show that news stories contain information relevant to predicting both earnings and stock returns
  - He analyzed the Wall Street Journal's "Abreast of the Market" column, and Tetlock et al. (2008) extended the analysis to company specific stories in the Wall Street Journal and the Dow Jones News Service
  - Subsequent studies applied similar techniques with a variety of sources
- Researchers have found that textual information from news can briefly predict returns at the aggregate and also at the individual stock level
- However, most research has been limited to a narrow event window within two days after the news release

Heston and Sinha (2017, FAJ)

- The duration and reversal of return predictability are important because they affect the economic interpretation of news in terms of permanent news impact or transient sentiment
- Sentiment theory predicts short-horizon returns will be reversed in the long run, vs. information theory predicts they will persist
- Heston and Sinha (2017) explore the temporal pattern of predictability in individual stock returns using a sophisticated neural network
- They apply these techniques to 900,754 articles tagged with company identifiers from the Thomson Reuters NewsScope system over the years 2003–2010, not only to newspapers or newswire services
  - Their data include a measure of the “tone” or sentiment of each story
  - The algorithms synthetically read news in 3 steps, (i) preprocessing, (ii) lexical and sentiment pattern identification, and (iii) classification
  - The first two stages of the sentiment engine identify parts of speech and morphologically stem the words by matching each word to its root word
  - The engine performs parsing by identifying the subject of the sentence and identifies words as adjectives, adverbs, intensifiers, nouns, and verbs; the lexical identification also recognizes negation and intensification

Heston and Sinha (2017, FAJ)

- The neural network computes an intermediate “hidden” layer, which connects to the final output classification layer
- The weights of connections within the NN are chosen to optimize the accuracy of prediction compared with human classification of articles
- The classifier was trained using a random sample of 3,000 triple-annotated news articles spanning the 14 months from December 2004 to January 2006
- Companies in the largest decile had frequent stories—22.42 stories per week, which exceeds 3 stories per day—but the small companies were comparatively neglected in news coverage
- They find that the neural network appears to extract permanent information that is not fully incorporated into current stock prices
- Stocks with positive (negative) news on one day have predictably high (low) returns for the subsequent one to two days
  - The most dramatic difference occurred in the smallest decile, where the average small company lost 0.14% in a given week but small companies with news averaged +2.00% in the week following the news

Heston and Sinha (2017, FAJ)

- o Avg. excess returns on the quintile spreads are invariably positive in the 10 days preceding the publication of news
- o Returns predict news, rather than converse

- Aggregating news over one week produces a dramatic increase in predictability of returns: a decile spread of stocks based on news over the past week earns average excess returns exceeding 2% over the subsequent 13 weeks
- Positive news predicts positive returns for only about one week, but negative news predicts negative returns for up to a quarter
- Reaction to negative news over longer horizons suggests that short-sale constraints might slow the incorporation of information extracted by our textual processing techniques

Table 3. Long–Short Excess Return Based on News Sentiment on Day 0

	Thomson Reuters Quintiles	
Day after News	Mean	*t*-Statistic
−9	0.09%	6.0
−8	0.07	4.5
−7	0.09	5.9
−6	0.10	6.3
−5	0.12	7.4
−4	0.08	4.7
−3	0.12	6.9
−2	0.18	10.8
−1	0.50	22.4
0	1.99	63.9
1	0.17	9.8
2	0.04	2.5
3	0.02	1.2
4	0.04	2.5
5	0.03	1.6
6	0.06	0.4
7	0.02	1.1
8	0.01	0.9
9	−0.02	−1.2
10	−0.00	0.0

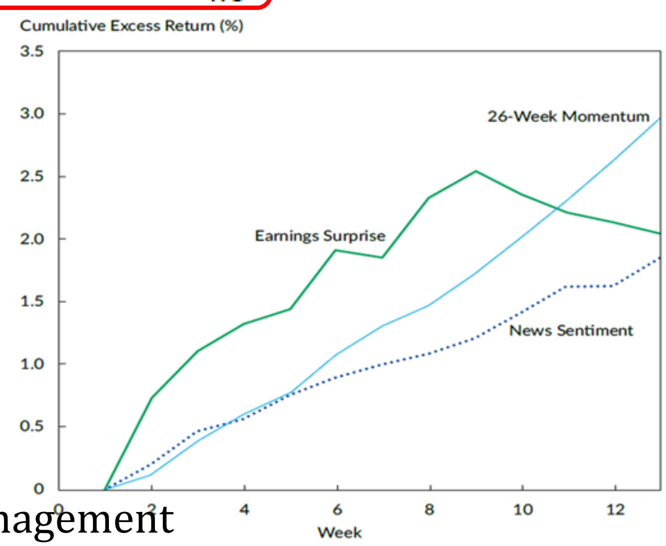
Heston and Sinha (2017, FAJ)

Table 4. Weekly Returns from Long–Short Portfolio Based on News in Week 0

Week after News	Return	*t*-Statistic	Momentum-Adjusted Return	*t*-Statistic	Size-Adjusted Return	*t*-Statistic
A. Long–short excess returns from weekly portfolio for all stocks with news						
0	3.75%	37.0	3.61%	41.9	3.62%	36.8
1	0.32	3.9	0.31	2.5	0.36	2.5
2	0.20	2.6	0.12	1.0	0.20	1.4
3	0.26	3.6	0.22	1.7	0.25	1.8
4	0.10	1.4	0.01	0.1	0.02	0.1
5	0.19	2.6	0.13	1.0	0.21	1.5
6	0.14	1.9	0.22	1.8	0.25	1.9
7	0.11	1.5	0.00	0.0	0.02	0.2
8	0.08	1.2	0.14	1.1	0.19	1.4
9	0.12	1.6	0.23	1.8	0.24	1.8
10	0.21	2.8	0.23	1.6	0.22	1.5
11	0.20	2.9	0.29	2.3	0.36	2.6
12	0.01	0.2	0.05	0.4	0.06	0.5
13	0.21	2.6	0.27	2.1	0.27	2.0
Weeks 1–13	2.15	8.2	2.22	4.8	2.65	5.3
Weeks 2–13	1.83	7.4	1.91	4.3	2.29	4.8

Sentiment spread returns are not explained by size or momentum

- The cumulative returns from the weekly news strategy show a persistent upward trend
- Two explanations for the improvement in predictability when using weekly returns: (i) some companies have multiple news stories over different days within a week



Heston and Sinha (2017, FAJ)

- (ii) the distribution of daily news is quite variable over time

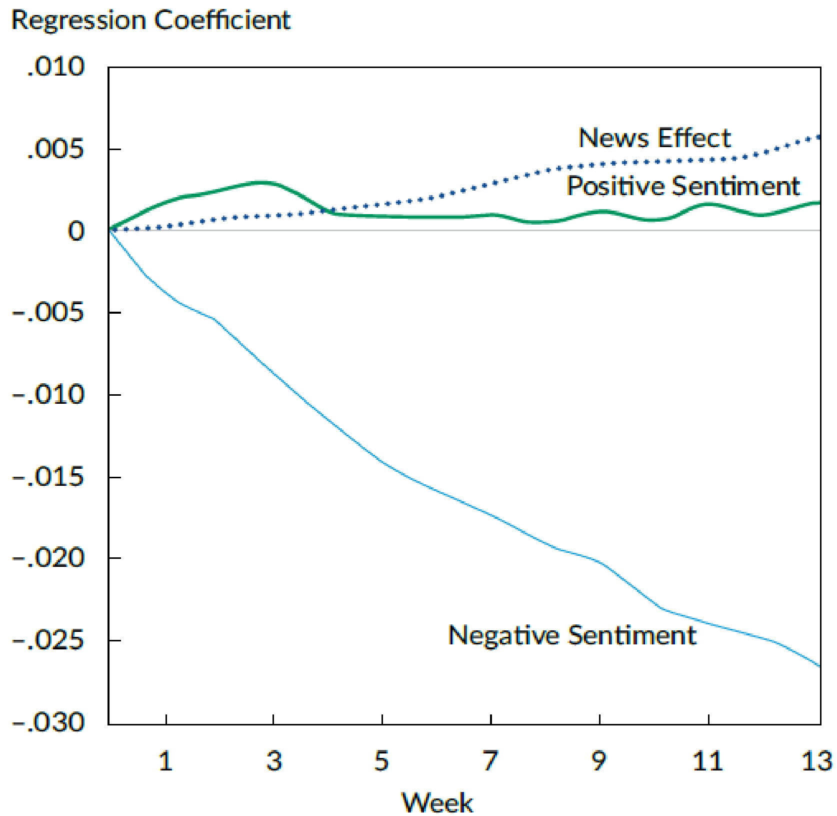
$$r_{i,t} = \alpha_{k,t} + \gamma_{k,t}\text{If\_news}_{i,t-k} + \beta_{k,t}\text{Positive}_{i,t-k} + \delta_{k,t}\text{Negative}_{i,t-k} + \varepsilon_{i,t}$$

Table 6. Cross-Sectional Regressions of Weekly Returns Based on Thomson Reuters Sentiment in Week 0: Average Cross-Sectional Regression Coefficient on Thomson Reuters Sentiment Variables

Week after News	$\mathcal{D}$	*t*-Statistic	News Effect	*t*-Statistic	Positive Sentiment	*t*-Statistic	Negative Sentiment	*t*-Statistic
0	−0.0005	−0.3	−0.0003	−0.4	0.0304	27.6	−0.0329	−23.5
1	−0.0005	−0.3	0.0008	1.4	0.0017	2.0	−0.0037	−3.7
2	−0.0006	−0.4	0.0008	1.5	0.0008	0.9	−0.0018	−1.9
3	−0.0006	−0.4	0.0012	2.2	0.0004	0.5	−0.0032	−3.5
4	−0.0005	−0.3	0.0019	3.4	−0.0016	−1.9	−0.0027	−2.7
5	−0.0004	−0.3	0.0012	2.2	−0.0003	−0.4	−0.0026	−2.9
6	−0.0003	−0.2	0.0009	1.5	−0.0002	−0.2	−0.0018	−1.9
7	−0.0003	−0.2	0.0010	1.8	0.0002	0.2	−0.0015	−1.6
8	−0.0003	−0.2	0.0012	2.1	−0.0005	−0.7	−0.0018	−1.9
9	−0.0002	−0.2	0.0004	0.7	0.0007	0.9	−0.0011	−1.1
10	−0.0001	0.0	0.0011	2.0	−0.0005	−0.7	−0.0026	−2.8
11	−0.0004	−0.3	0.0003	0.6	0.0009	1.2	−0.0011	−1.3
12	−0.0004	−0.3	0.0013	2.4	−0.0005	−0.7	−0.0009	−1.0
13	−0.0002	−0.2	0.0004	0.8	0.0008	1.0	−0.0017	−1.8
Total					0.0321		−0.0593	

Bad (good) news travel slowly (fast)

Heston and Sinha (2017, FAJ)



Note: This figure plots the cumulative coefficients from Table 6 for horizons ranging from 1 week to 13 weeks.

- In addition to the traditional news sources, there is now a plethora of social media through which views, ideas, and opinions are exchanged
    - As of August 2018, 68% of Americans report that they get at least some of their news on social media—with 2 in 10 doing so often
- Social media are unstructured and opinionated and have unfiltered content

Beckers (2019, FAJ)

- Academic studies of social media–based sentiment mainly use Twitter data (besides Yahoo Finance, Raging Bull, and Seeking Alpha)
- The majority of these studies rely on a history of less than a year to check for short-term (daily) forecasting ability: the short history and limited sample make the inferences of these papers hard to generalize
- Beckers (2019) studies whether news or social media contain any information that is of relevance for investment decision making and if so whether the two sources are complementary or overlapping
- He uses Thomson Reuters MarketPsych Indices (TRMI) for which the news media articles are sourced from more than 2,000 global news publications with the highest rank by external links (a measure of credibility and readership)
    - The social media content derives from stock forums, tweets, comment streams, and blogs and includes over 700 primary sources
- Dating back to 1998:3 TRMI has analyzed billions of social media and news articles that are referenced and mapped to over 8,000 equities, 52 equity indexes, 32 currencies, 35 commodities, and 130 countries

Beckers (2019, FAJ)

- Beckers focusses on three types of TRMI data: (i) emotional indexes, (ii) fundamental equity information, (iii) general political/macro environment reflecting trust, conflict, and time urgency
- He reconstructs a global indicator for emotion, equity fundamentals, and political risk by calculating an appropriately cap-weighted monthly average of the constituent country indexes from March 1998 through December 2017

Aggregate Sentiment Indices and their TRMI Constituent Parts

Index	TRMI Contributing Indexes	Description	Sign
Emotion	Sentiment	Overall positive references, net of negative references	+
	Optimism	Optimism, net of references to pessimism	+
	Fear	Fear and anxiety	−
	Joy	Happiness and affection	+
	Gloom	Gloom and negative future outlook	−
	Stress	Distress and anger	−
	Love	Love net of references to hate	+
	Anger	Anger and disgust	−
Equity Fundamentals	Long–Short	Net buying/selling	+
	Long–Short Forecast	Forecasted net buying/selling	+
	Price Direction	Net price increases	+
	Price Forecast	Forecasted net price increases	+
	Debt Default	Debt defaults and bankruptcies	−
	Innovation	Innovativeness	+
	Analyst Rating	Net analyst upgrades	+
	Dividends	Net dividend increases	+
	Earnings Forecasts	Net earnings changes	+
	Fundamental Strength	Net positivity about accounting fundamentals	+
	Layoffs	Net staff reductions	−
	Litigation	Litigation and legal problems	−
	Management Trust	Trust in management team, net of references to unethical behavior	+
	Mergers	Mergers or acquisitions	+
Political Risk	Trust	Trustworthiness, net of references to corruption	+
	Violence	Violence, social unrest and war	−
	Conflict	Disagreement	−
	Time Urgency	Timeliness and urgency net of references to lethargy and delays	+

Beckers (2019, FAJ)

Correlations of Monthly Changes, March 1998 to December 2017

			Month t					
			Social Media			News		
			Emotion	Equity Fund	Pol Risk	Emotion	Equity Fund	Pol Risk
Month t	Social Media	Emotion	1.00					
		Equity Fund	0.47	1.00				
		Pol Risk	0.40	0.45	1.00			
	News	Emotion	0.41	0.37	0.12	1.00		
		Equity Fund	0.23	0.40	0.13	0.40	1.00	
		Pol Risk	0.10	0.23	0.36	0.11	0.35	1.00
Month $t-1$	Social Media	Emotion	–0.26	0.07	–0.05	–0.05	–0.08	–0.01
		Equity Fund	–0.15	–0.11	–0.07	–0.07	–0.14	–0.02
		Pol Risk	–0.03	–0.02	–0.13	0.06	–0.02	–0.05
	News	Emotion	–0.14	0.07	–0.05	–0.19	–0.22	–0.08
		Equity Fund	–0.05	0.05	–0.03	0.05	–0.39	–0.13
		Pol Risk	–0.02	–0.08	–0.15	0.11	–0.13	–0.38

- The contemporaneous cross correlations are all respectably high, indicating that all variables essentially capture similar information
- Beckers correlates series with next month's MSCI World index return
- Newsbased variables have on average promising correlations with future returns

	Emotion Social Media	Equity Fund Social Media	Political Risk Social Media	Emotion News	Equity Fund News	Political Risk News	Emotion News + SM	Equity Fund News + SM	Political Risk News + SM
1998–2002	–0.12	–0.14	–0.04	–0.31	0.14	–0.02	–0.26	0.02	–0.02
2003–2007	0.04	–0.09	–0.01	0.04	0.08	–0.03	0.01	0.05	0.02
2008–2012	0.15	0.17	–0.07	0.20	0.09	0.00	0.20	0.17	–0.04
2013–2017	–0.05	–0.10	0.06	0.12	0.00	0.12	0.01	–0.01	0.13
1998–2007	–0.09	–0.09	–0.02	–0.21	0.12	–0.02	–0.18	0.04	–0.01
2008–2017	0.07	0.11	–0.02	0.18	0.07	0.03	0.15	0.11	0.00
1998–2017	0.00	0.02	–0.02	–0.03	0.09	0.00	–0.02	0.07	0.00

Beckers (2019, FAJ)

- Social media–based signals have patchy predictive power and tend to be dominated by the corresponding news-based variables
- An (equally weighted) combination of news and social media sentiments does not materially improve on the pure news-based signals (except for the equity fundamentals during the 2008–2012 period)
- He also investigates an investment strategy
  - If the monthly change in any sentiment variable exceeds 10% (in absolute value), we assume we invest 130% (70%) in next month's index
  - If the change is less than ±10%, we are just 100% long the market
- This strategy yields a active return (strategy – market), of which we calculate the annualized information ratio (IR), see Appendix
- The news-based IRs are almost consistently higher, and the combined signals produce less profitable results than the one solely news based
- A social media political risk–based strategy does, however, seem to yield more reliable returns than one based on news sources
- The importance of the social media signals does not grow with the proliferation of the social media sources through time

Beckers (2019, FAJ)

Information Ratio of Market Timing Strategy

	Emotion			Equity Fundamentals			Political Risk						
	Social Media	News	Social Media + News	Social Media	News	Social Media + News	Social Media	News	Social Media + News	Average all SM	Average all News	Average all SM + News	SM Pol Risk + Fund Eq News
1999	−0.15	−1.09	−0.66	−0.27	−2.44	−1.88	0.48	−1.90	−1.37	−0.03	−2.35	−1.48	−1.74
2000	0.15	1.06	0.69	0.56	0.61	0.63	0.09	−1.05	−0.62	0.31	0.18	0.27	0.34
2001	−0.06	−1.33	−0.77	0.14	−0.74	−0.33	0.09	0.24	0.17	0.06	−0.73	−0.33	−0.35
2002	−1.67	−1.38	−1.63	−0.98	0.73	−0.09	−0.74	−0.33	−0.56	−1.18	−0.42	−0.87	−0.02
2003	0.04	0.89	0.56	−0.54	0.27	−0.15	−0.43	0.06	−0.18	−0.36	0.48	0.07	−0.10
2004	0.21	0.94	0.64	0.44	1.21	1.11	0.99	0.76	0.95	0.64	1.18	1.03	1.16
2005	1.13	1.18	1.37	2.02	−0.35	0.84	2.05	0.78	1.60	2.42	0.65	1.73	1.03
2006	0.70	0.55	0.73	0.10	−0.35	−0.13	1.26	−1.04	0.15	0.79	−0.49	0.34	0.77
2007	−1.29	0.51	−0.42	−0.74	−0.22	−0.52	−2.31	−1.77	−2.11	−1.51	−0.52	−1.09	−1.36
2008	−0.02	−0.45	−0.23	0.64	−0.28	0.17	2.06	0.90	2.53	1.02	0.18	0.78	0.99
2009	0.72	2.37	1.96	0.59	1.96	1.38	−2.11	0.35	−0.94	−0.40	1.98	0.92	−0.10
2010	−0.78	−0.24	−0.55	−0.91	−0.59	−0.87	−0.88	−0.60	−1.25	−0.95	−0.70	−1.08	−0.89
2011	−0.74	0.22	−0.49	−0.96	0.07	−0.44	−0.48	−1.15	−1.13	−0.77	−0.56	−0.76	−0.24
2012	1.41	0.51	1.09	0.78	1.36	1.11	−1.44	−0.69	−1.31	0.36	0.76	0.56	−0.06
2013	0.00	0.36	0.14	0.06	−1.40	−0.53	0.39	−1.09	−0.26	0.15	−0.98	−0.22	−0.26
2014	−1.13	−1.59	−1.72	0.63	0.38	0.65	1.74	0.72	1.56	0.56	−0.27	0.29	1.67
2015	0.07	1.60	1.11	−1.46	0.90	−0.81	0.70	0.36	0.54	−0.21	1.45	0.57	0.89
2016	0.21	−0.55	−0.31	−0.05	0.06	0.01	2.25	0.28	2.32	1.00	−0.05	1.09	2.73
2017	−0.65	0.13	−0.29	0.38	0.70	0.62	−0.11	0.55	0.26	−0.11	0.66	0.29	0.31
1998–2002	−0.62	−0.90	−0.84	−0.31	0.03	−0.19	−0.29	−0.40	−0.40	−0.45	−0.57	−0.57	−0.16
2003–2007	0.08	0.76	0.50	0.03	0.16	0.11	0.17	−0.22	−0.03	0.11	0.28	0.21	0.20
2008–2012	0.06	0.31	0.20	0.10	0.46	0.30	−0.40	−0.03	−0.31	−0.12	0.41	0.12	0.01
2013–2017	−0.17	0.18	−0.02	−0.17	0.05	−0.09	0.92	0.17	0.67	0.21	0.19	0.28	0.79
1998–2007	−0.35	−0.23	−0.32	−0.18	0.08	−0.07	−0.10	−0.32	−0.24	−0.23	−0.21	−0.25	−0.01
2008–2017	−0.02	0.25	0.13	0.01	0.32	0.18	−0.01	0.03	0.01	−0.01	0.32	0.15	0.19
1998–2017	−0.17	0.02	−0.08	−0.08	0.21	0.08	−0.05	−0.13	−0.11	−0.12	0.05	−0.05	0.10
Yearly IR > 0 1998–2017	10	12	9	11	11	9	11	10	9	10	9	12	9
Yearly IR > 0 2008–2017	5	6	4	6	7	6	5	6	5	5	5	7	5

Appendix: The Information Ratio

Goodwin's (1997) *information ratio* (also known as an signal-to-noise ratio). This statistic measures a portfolio's average return in excess of that of a benchmark portfolio divided by the standard deviation of this excess return, for instance:

$$\boxed{T^{-1}\sum_{t=1}^{T}(R_{FMV,t}-R^{f})-\hat{\beta}_{FMV}\left[T^{-1}\sum_{t=1}^{T}(R_{t}^{m}-R^{f})\right]}$$

$$IR_{FMV} \equiv \frac{\hat{\alpha}_{FMV}^{JENSEN}}{\sqrt{T^{-1}\sum_{t=1}^{T}[(R_{FMV,t}-R^{f})-\hat{\beta}_{FMV}(R_{t}^{m}-R^{f})]^{2}}}$$

Sometimes the numerator is simply stated as the difference between the sample mean excess return on a portfolio and the same quantity for some appropriate benchmark. For instance, the numerator in equation (7.8) may be replaced with $T^{-1}\sum_{t=1}^{T}(R_{FMV,t}-R^{f})-T^{-1}\sum_{t=1}^{T}(R_{t}^{m}-R^{f})$, when the market portfolio represents a sensible benchmark (for instance, for an equity index fund that typically shows a correlation with large, aggregate stock indices in excess of 0.98). In this case, the standard error estimate that appears at the denominator, $\left[T^{-1}\sum_{t=1}^{T}(R_{FMV,t}-R_{t}^{m})^{2}\right]^{1/2}$ is called the *tracking error* of the FMV. More generally, the information ratio is the ratio between the mean return of a portfolio in excess of a given benchmark divided by the standard deviation of such excess return.