Big Data and AI Strategies
——Machine Learning and **Alternative Data** Approach to Investing

J.P. Morgan, 2017. 5

吕漫妮

2020. 11. 14

Contents

Backgrounds

- Big Data Times: Given the amount of data that is available, a skilled quantitative investor can nowadays in theory have near real time macro or company specific data not available from traditional data sources.
- Driving factors:

① Exponential increase in amount of data available

② Increase in computing power and data storage capacity, at reduced cost

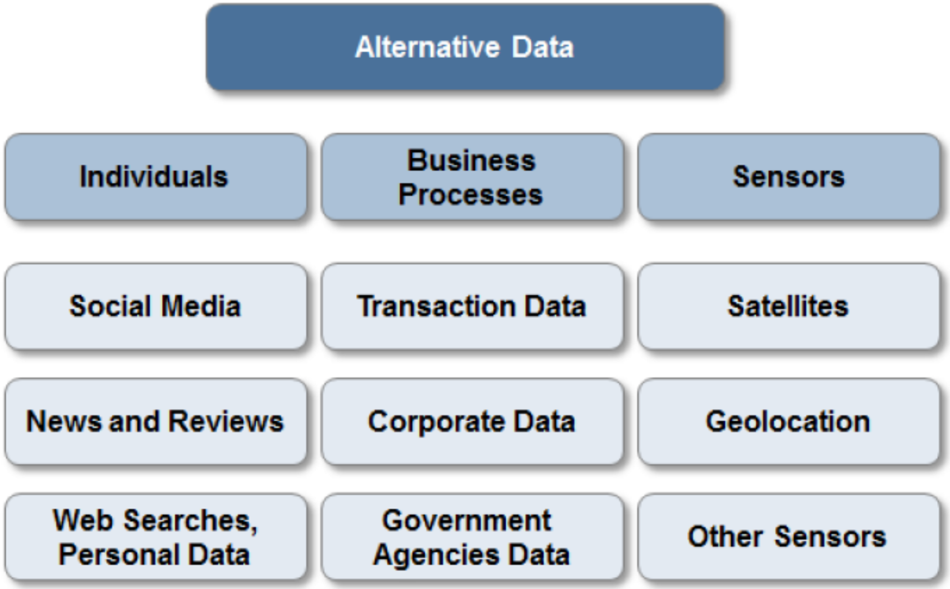
③ Advancement in Machine Learning methods to analyze complex datasets

- Necessity: A new source of competitive advantage is emerging with the availability of alternative data sources as well as the application of new quantitative techniques of Machine Learning to analyze these data. The market will start reacting faster and will increasingly anticipate traditional or 'old' data sources.

Alternative datasets

- Definition: non-traditional data that can be used in the investment process
- Features: often larger in volume, velocity and variability as compared to traditional datasets
- Alternative datasets types:

① Classifying data based on the manner in which the data was generated



Alternative datasets

- Alternative datasets types:

② We consider attributes that are directly relevant for investment professionals ('Investment Classification')

Asset Class	Investment Style	Alpha (Net of Cost)	Known	Stage of Processing	Quality	Technical Aspects
Equity	Macro	Viable Stand alone	Public Free of cost	Raw	History	Frequency
Commodity	Sector Specific	Viable In a Portfolio	Well Known	Semi Processed	Outliers	Latency
Credit	Stock Specific	Not Viable	Lesser Known	Processed	Missing Values	Format
Rates	Risk Indicator	Capacity	Proprietary Not Known	Trading Signal	Methodology Transparency	Robust API
FX	Quant Signal	Orthogonality	Limited Sales Deals	Research Piece or Alert	Support Structure	Conflicts and Legal Risk

CIOs and Portfolio Managers ⟷ Quants and Data Scientists

- Big Data Compliance

Alternative datasets

- Potential Pitfalls:

① Datasets may don't contain alpha, signals that have too little investment capacity, decay quickly, or are simply too expensive to purchase relative to their benefit.

② Managers may invest too much into unnecessary infrastructure that don't justify marginal performance improvements.

- Three types of data providers in the marketplace:

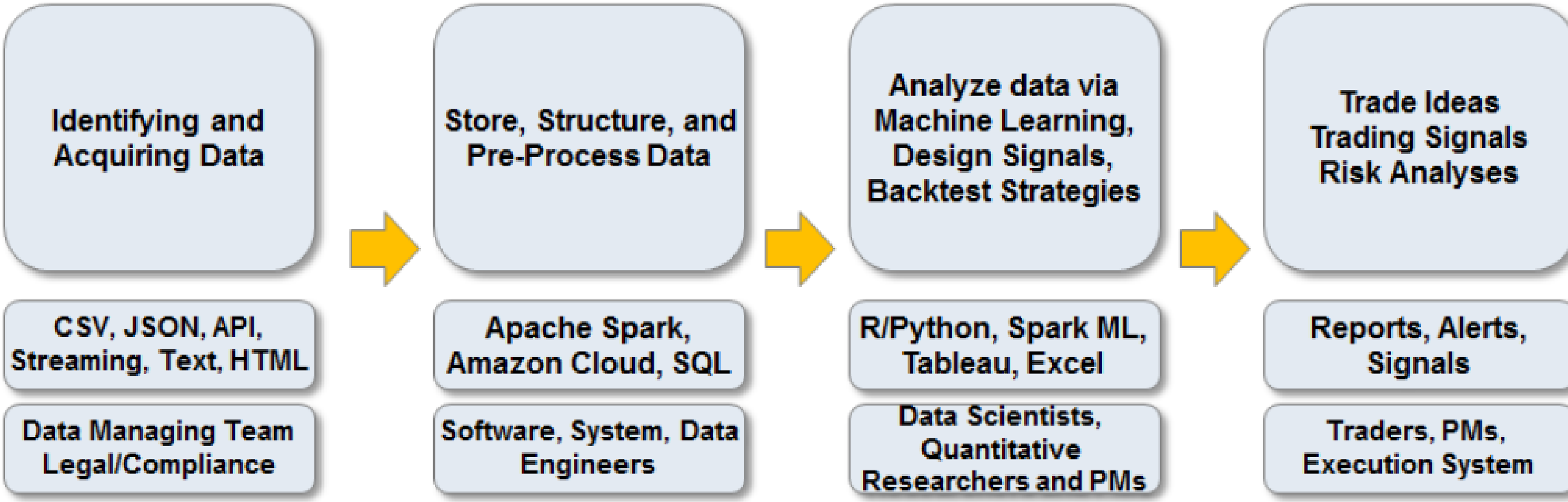
① Providers of raw data collect and report alternative data with minimal aggregation or processing.

② Providers of semi-processed data partially process data by aggregating over geographic regions, industry sectors or map Big Data to specific securities.

③ Providers of signals and reports are focused on the investment industry alone.

Alternative datasets

- Big Data workflow for investment:



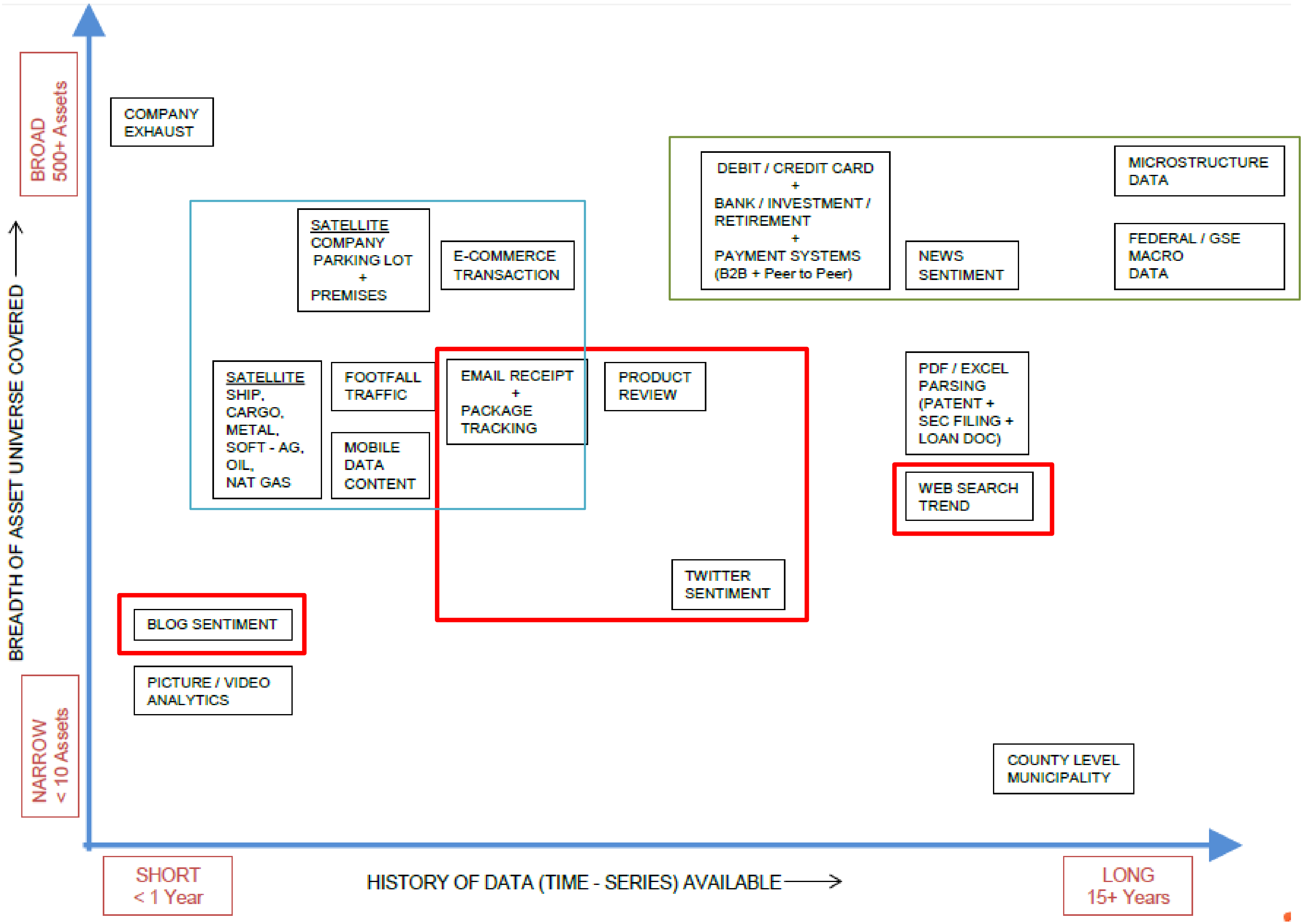
- Types:

① Data from Individual Activity

② Data from Business Processes

③ Data from Sensors

Typical length of history for alternative data sets



Alternative datasets

- Data from Individual Activity

① social media data (e.g. Twitter, LinkedIn, blogs) - Social media sentiment analysis is fairly popular.

② data from specialized sites (e.g. news media, product, reviews)

③ web searches and volunteered personal data (e.g. Google search, email receipts).

Case Study: Using Twitter Sentiment to trade S&P 500 (iSentium)

Case Study: Using News Sentiment to trade Bonds, Currencies and Commodities (Ravenpack)

- Data from Business Processes

① data from public agencies (e.g. federal and state governments)

② data from commercial transactions (including e-commerce and credit card spending, exchange transaction data)

③ data from other private agencies (e.g. industry specific supply chain data).

Case Study: Using Email Receipt Data to trade US Equities (Eagle-Alpha)

Alternative datasets

- Data from Sensors

① satellite data

② geolocation data

③ data generated by other sensors

Case Study: Using Cellular Location to Estimate Retail Sales (Advan)

Case Study: Satellite Imagery of Parking Lots and Trading Retail Stocks (RS Metrics)

How news and its context drive risk and returns around the world

Charles W. Calomiris, Harry Mamaysky
The Journal of Finance, 2019.8

吕漫妮
2020. 11. 14

Contents

Backgrounds & Motivation

- What is news and how is it associated with changes in stock market returns and risks? This is a fundamental question in asset pricing and has been the subject of decades of research.

➢ The promising early work in the literature linking textual analysis and stock returns has raised more questions that it has answered.

Research Problem

- This paper addresses nine important sets of questions about the connections between news and market outcomes.

1. How should one best measure news using word flow?
   - One approach is to apply methods with no a priori position regarding which particular words should be the focus of the analysis to organize the flow of words in a comprehensive and unconstrained manner to see which parts of word flow matter.

     Another approach is to identify based on a priori lists of words.
2. Which aspects of word flow should be the focus of measurement?
   - In addition to measuring sentiment, the contextual frequency of word flow, and the way sentiment matters differently depending on context, other aspects of text flow may be relevant.

Research Problem

- This paper addresses nine important sets of questions about the connections between news and market outcomes.

3. The patterns that link frequency, topics, sentiment, and entropy measures of word flow with market outcomes may vary over time.
   - We capture changes over time using a dividing point that is identified by principal components analysis. We further explore dynamic changes in coefficients using a rolling elastic net regression.
4. Given the potential importance of identifying topical context, how should one identify topics?
   - Within the set of non-priori means of identifying topics there are two common methods, namely the Louvain and latent Dirichlet allocation (LDA) approaches, as discussed below.

Research Problem

- This paper addresses nine important sets of questions about the connections between news and market outcomes.

5. Does the effect of our measures operate through a risk channel?
    - Our findings suggest that when a word flow measure predicts positive expected returns, it also predicts a reduction in risk. This opposite effects fact that news tends to have suggests that the factors captured by news flow are not priced risks.
6. How should one measure risk? As is well known, if the returns process is characterized by Brownian motion and normality of the error term, then the standard deviation of returns will be a sufficient statistic for risk.
    - In addition to using the standard deviation of returns (sigma), we also employ the "maximum one-year drawdown."

Research Problem

- This paper addresses nine important sets of questions about the connections between news and market outcomes.

7. The existing literature focuses on short-term analysis of individual US companies or the US stock market. Do empirical patterns that apply to individual company stocks or the aggregate U.S. index also apply to other countries?
    - Because returns processes, the amount of risk and the nature of the news that drives risk differ between emerging markets (EMs) and developed markets (DMs), we divide countries into EMs and DMs and perform separate panel analyses of each group of countries.
8. What source of news should one use?
    - Thomson Reuters's entire database of news articles from 1996 through 2015——an English language news source covering many countries.

Research Problem

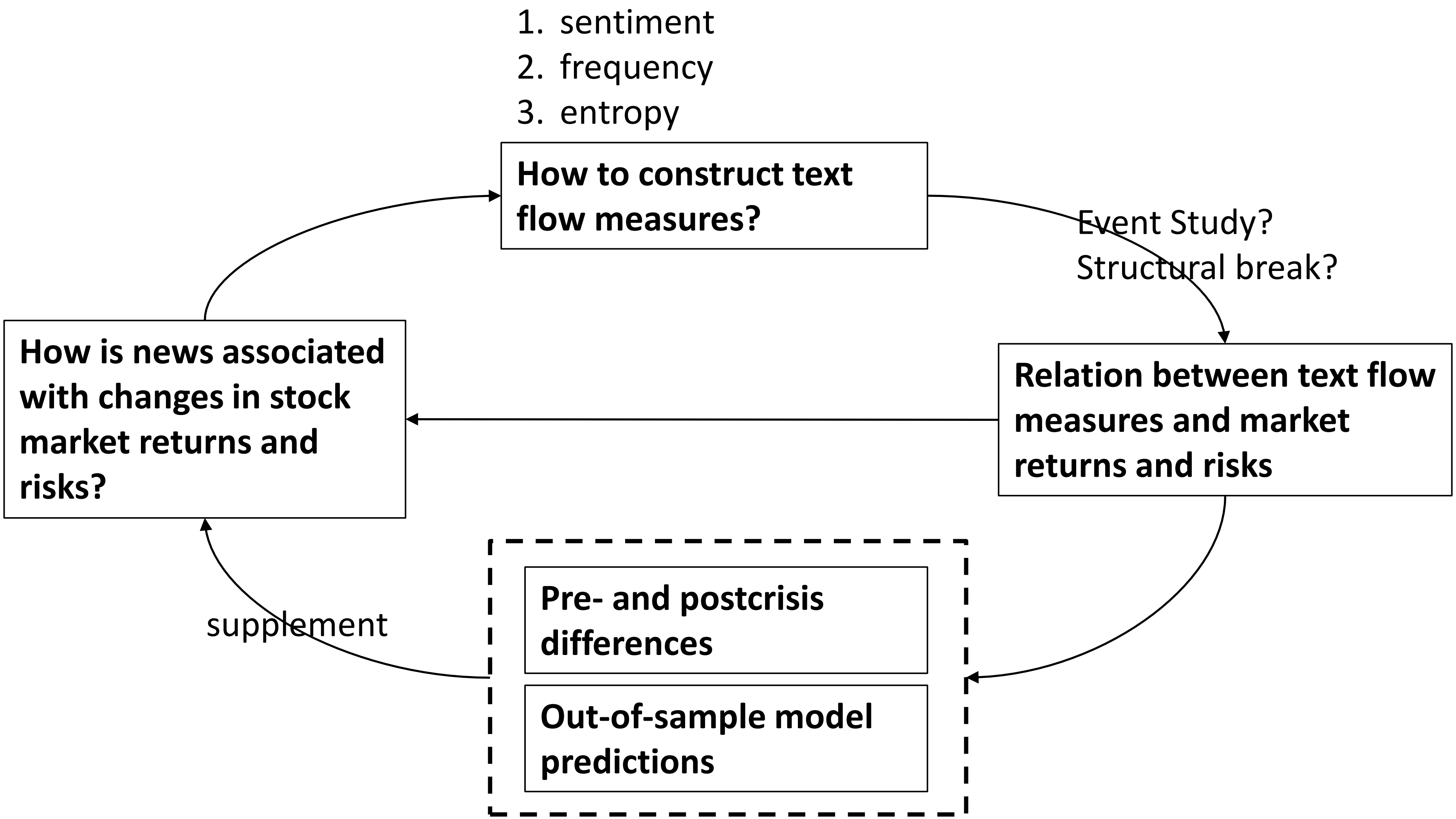
- This paper addresses nine important sets of questions about the connections between news and market outcomes.

9. Over what time frame should word flow predict risk and return?
    - We aggregate news at a monthly horizon, examine both one-month-ahead and one-year-ahead predictions, and show that our country-level measures exhibit stronger predictive power for one-year-ahead returns and drawdowns.

Contribution

- This paper develops an atheoretical approach for capturing news through various word flow measures and apply that approach to 51 countries over the time period 1998–2015.
- This paper gives out a trading strategy based on out-of-sample model predictions, using an elastic net regression.

Outline


1. sentiment
2. frequency
3. entropy
How to construct text flow measures?
Event Study?
Structural break?
How is news associated with changes in stock market returns and risks?
Relation between text flow measures and market returns and risks
supplement
Pre- and postcrisis differences
Out-of-sample model predictions

Model Design: Data

- Data:

1. News: Our textual data source is the Thomson Reuters Machine Readable News archive, includes all the English language Reuters News articles from 1996 to 2015, covering a wide range of topics.
2. Country-level stock market index data: obtained from Bloomberg and are converted into US dollar terms using end-of-day exchange rates from Bloomberg.
3. Macro data (such as interest rates, GDP growth rates, and credit ratios): obtained from the World Bank and the International Monetary Fund etc.

Model Design: Measures and Construction

- Measures:

We use our own sentiment measures constructed directly from the text.

- Construction Steps

1. Corpus selection and cleaning
2. Construction of the document term matrix and topic classification
3. Extraction of 4-grams to allow for calculation of entropy measures
4. Calculation of article-level sentiment, topic measures

Model Design: Construction of measures

- Construction Steps

1. Corpus selection and cleaning

Our EM corpus consists of all articles tagged by Thomson Reuters with the *N2:EMRG* code. Our DM corpus consists of all articles about the countries identified as developed market economies.

All textual analysis in the paper is done separately for the EM and DM corpora.

2. Construction of the document term matrix and topic classification

Document term matrix: rows→article, columns→words in our *econ word* list, counts the number of times a given word appears

econ word list is a list of 1242 stemmed words, bigrams, and trigrams that are descriptive of either market or economic phenomena.

Model Design: Construction of measures

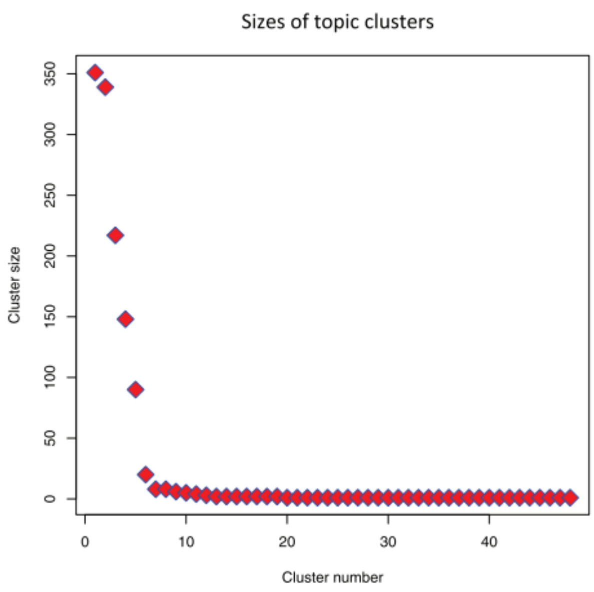
2. construction of the document term matrix and topic classification

We look for non-overlapping clusters (groups of words) of that tend to occur together in articles frequently.

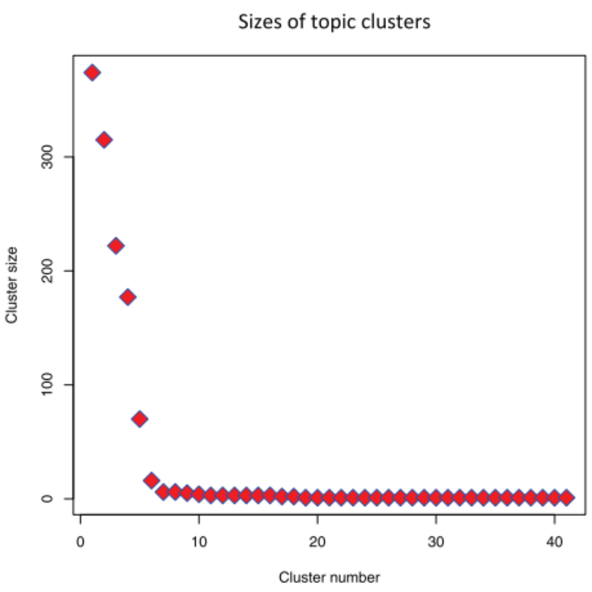
① Format the cosine similarity matrix

② maximize network modularity via the Louvain algorithm

Developed markets



Emerging markets



Model Design: Construction of measures

- Construction Steps

2. Construction of the document term matrix and topic classification

① For the EM corpus, we find five word-groups: markets (Mkt), governments (Govt), commodities (Comms), corporate governance and structure (Corp), and macroeconomic topics (Macro).

② For the DM corpus, we find the first four topics and the extension of credit (Credit).

The word overlap between the topics in our EM and DM corpora is often sizable.

Each cluster shows the number of occurrences (in millions) of its constituent words in the corpus.

Mkt # words (mm) = 236.23

Model Design: Construction of measures

3. Extraction of 4-grams to allow for calculation of entropy measures

We calculate the conditional probability of observing the fourth word in the phrase conditional on seeing the first three words.

$$m = \frac{\hat{C}(w_1, w_2, w_3, w_4) + 1}{\hat{C}(w_1, w_2, w_3) + 10}$$

We extend the concept of entropy at the 4-g level to the article level by calculating the negative average log probability of all 4-g in article.

$$H_j = -\sum_i p_{j \cdot i} \log m_i$$

$p_{j \cdot i}$ is the fraction of all 4-g appearing in article j represented by the i^{th} 4-g.

We characterize an article as unusual if it contains language that is unlikely to have been seen in the past, which may indicate heightened or lower market risks.

Model Design: Construction of measures

- Construction Steps

4. calculation of article-level sentiment, topic measures

We use the Loughran McDonald (2011) sentiment word lists to calculate article-level sentiment measure, for article j

$$s_j = \frac{POS_j - NEG_j}{a_j}$$

For topic τ, we define $e_{\tau,j}$ as the number of econ words in article j that fall into topic τ, and e_j as the total number of econ words in article j, then $f_{\tau,j} = e_{\tau,j}/e_j$ represents the fraction of article j 's econ words that fall into a specific topic.

Model Design: Construction of measures

4. calculation of article-level sentiment, topic measures

We decompose an article's sentiment into a context-specific sentiment measure via

$$s_{\tau,j} = f_{\tau,j} \times s_j$$

We compute article-level context-specific sentiment interacted with entropy

$$SentEnt_{\tau,j} = f_{\tau,j} \times H_j \times s_j$$

which differentiates between topic sentiment on usual or unusual news days.

Model Design: Aggregating Measures

- Aggregation of article data at the daily and monthly level

The daily topic sentiment is

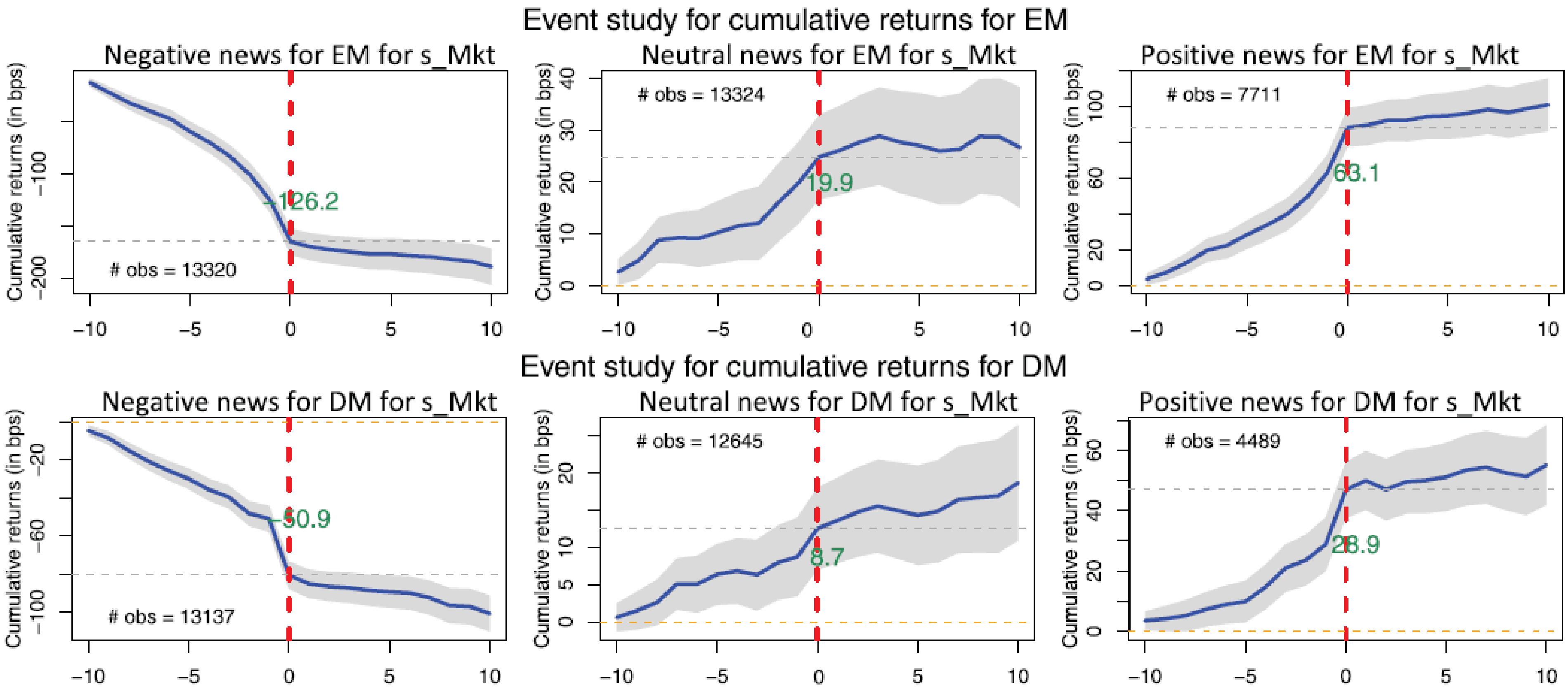
$$s_\tau = \sum_j \frac{a_j}{a} \times s_{\tau,j}$$

where a is the total number of words in all articles mentioning a given country on a given day.

The analogous definition is applied for article entropy and frequency.

The monthly measure for a given country is the simple average of that month's daily measures.

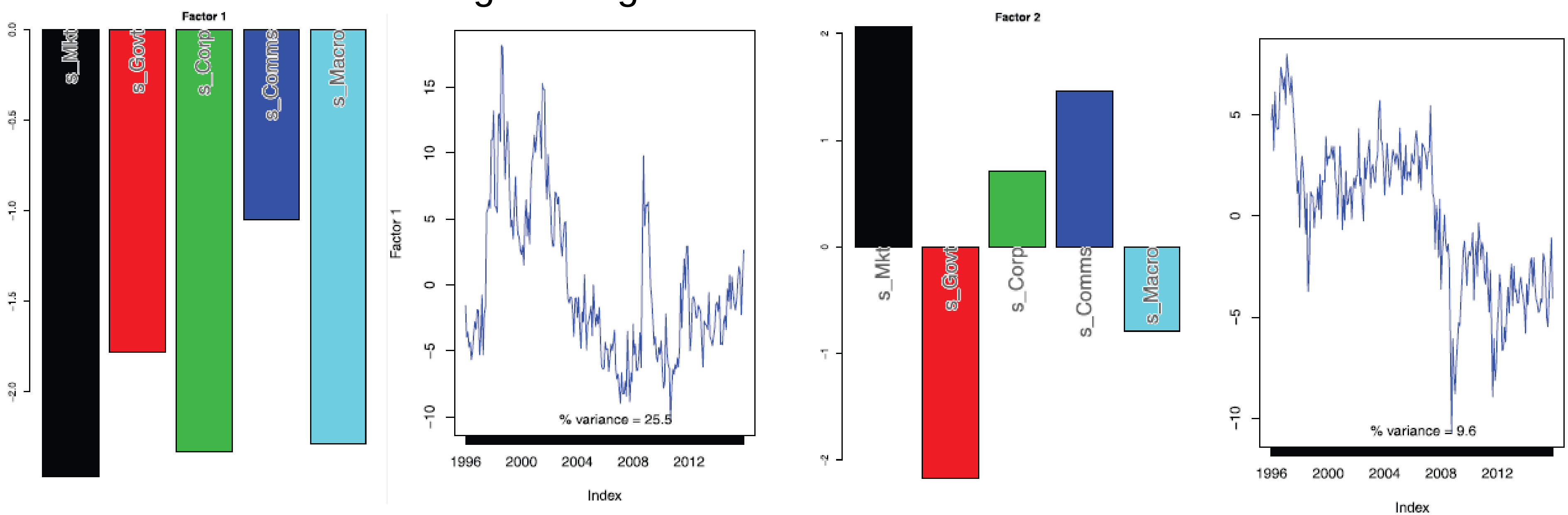
Empirical findings – Event Study



- The patterns are often similar. Cumulative returns tend to occur in advance of big news days, with the exception of negative news days for *Govt* and *Comms* in DMs and also positive news days for *Comms* in DMs.
- News events appear to cause more of a market reaction in our DM sample (more timely reporting in DMs or information leakage in EMs).

Empirical findings - Structural break around the financial crisis

- factor loadings and plots for each topic category related to the first two principal components for the time series of country-month-topic sentiment in DMs
- The first principal component tracks the aggregate time series of market sentiment. The second principal component appears as a step function with a break at the timing of the global financial crisis.



Each topic bar is the sum of that topic's factor loadings across all countries in the sample.

Empirical findings – Panel regressions for DMs next 12-month returns

topic sentiment measure in simple form

entropy interacted versions

Country-level monthly entropy

monthly average of daily article counts

s[Topic]: country level article sentiment

f[Topic]: frequency for Topic

Developed markets: Forecasting panel for next 12-month returns

	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	0.134	0.191	0.198	0.131	0.087	0.083	0.329*	0.372**	0.360**
$sigma_{t-2}$	0.037	0.076	0.079	-0.078	-0.011	-0.005	0.298	0.347*	0.332*
$return_{t-1}$	0.273	0.152	0.147	0.109	-0.026	-0.028	0.209	0.089	0.092
$return_{t-2}$	0.024	-0.070	-0.072	0.148	0.093	0.091	-0.110	-0.174	-0.173
$value_{t-1}$	21.240***	22.455***	22.563***	14.876***	13.254**	13.281**	25.244***	25.775***	25.869***
gdp_{t-1}	-0.473	-0.859	-0.904	0.232	0.269	0.224	-2.066***	-1.947***	-1.913***
$gdpdeflator_{t-1}$	0.878	0.486	0.465	0.580	0.499	0.460	0.056	0.119	0.126
cp_{t-1}	-0.276***	-0.220**	-0.224**	0.042	-0.016	-0.009	-0.301**	-0.292**	-0.298**
dcp_{t-1}	0.075	0.026	0.022	-0.177	-0.158	-0.158	0.270**	0.252*	0.257**
$rate_{t-1}$	-3.705***	-5.306***	-5.398***	-14.708***	-14.360***	-14.257***	-4.926***	-5.307***	-5.382***
$dexch_{t-1}$	0.392	0.418	0.414	-0.581	-0.705*	-0.711*	0.779	0.793	0.790
pre	-0.923	-2.502	-2.553	-3.088	-3.008	-3.003	2.835	2.010	2.102
$post$	-0.404	-1.365	-1.413	-2.708	-3.033	-2.968	2.283	1.991	2.067
$entropy_{t-1}$		20.794	23.123		-23.426	-21.945		23.165**	22.670*
$artcount_{t-1}$		-0.667	-0.709		0.379	0.299		-0.022	-0.008
$sMkt_{t-1}$		2.636	2.914		-1.356	-2.132		5.144*	5.637**
$fMkt_{t-1}$		0.755	0.864		-4.034	-4.135		1.245	1.231
$sGovt_{t-1}$		-3.268	-3.460*		-0.640	-0.965		-3.915*	-2.821
$fGovt_{t-1}$		-3.690	-3.840		-3.589	-3.809		-6.273	-6.006
$sCorp_{t-1}$		-2.767	-1.862		3.208	3.887*		-2.709	-2.885*
$fCorp_{t-1}$		-5.894*	-5.571*		-5.901	-5.760		-0.041	0.136
$sComms_{t-1}$		1.112	0.914		0.213	0.451		0.531	0.416
$fComms_{t-1}$		0.864	0.850		-1.537	-1.348		1.543	1.528
$sCredit_{t-1}$		3.764*	3.234		-1.007	-0.893		-0.743	-1.754
$fCredit_{t-1}$		0.092	0.084		0.853	0.832		-1.091	-1.295
R2	0.164	0.215	0.215	0.267	0.3	0.301	0.456	0.476	0.476
start	Apr 1998	May 1998	May 1998	Apr 1998	May 1998	May 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4411	4395	4395	2003	1987	1987	2408	2408	2408
stderr	both	both	both	both	both	both	both	both	both

Empirical findings – Panel regressions for EMs next 12-month returns

Emerging markets: Forecasting panel for next 12-month returns

	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	0.189	0.257**	0.256**	-0.045	0.255	0.254	0.486***	0.422***	0.417***
$sigma_{t-2}$	0.320**	0.373***	0.369***	-0.042	0.154	0.150	0.707***	0.662***	0.655***
$return_{t-1}$	0.298	0.199	0.206	0.091	-0.197	-0.184	0.440*	0.450**	0.464**
$return_{t-2}$	0.037	0.021	0.028	-0.099	-0.210	-0.200	0.111	0.131	0.142
$value_{t-1}$	4.061	4.075	4.181	3.657	8.939**	9.021**	9.371	7.417	7.521
gdp_{t-1}	-0.972	-0.884	-0.880	-0.849	-1.918*	-1.889*	-1.614*	-1.194*	-1.175
$gdpdeflator_{t-1}$	0.503*	0.387	0.389	1.078*	0.160	0.182	-0.267	-0.058	-0.061
cp_{t-1}	-0.487***	-0.311***	-0.307***	-0.320	-0.224	-0.212	-0.030	-0.125	-0.128
dcp_{t-1}	0.234	0.230	0.215	0.481	0.348	0.360	0.082	0.226	0.183
$rate_{t-1}$	-0.732*	-0.763*	-0.764*	-1.464*	-0.300	-0.317	-0.954	-1.146	-0.991
$dexch_{t-1}$	0.550	0.576	0.575	0.039	0.221	0.197	1.001**	0.811*	0.820*
pre	-0.277	-0.214	-0.203	-2.691	-5.186	-5.340	-1.580	-1.264	-1.258
$post$	-4.406	-4.149	-4.216	1.401	-0.147	-0.346	-7.895***	-7.018***	-7.167***
$entropy_{t-1}$		1.907	0.999		-48.869	-45.438		6.525	3.590
$artcount_{t-1}$		-5.785***	-5.788***		-10.478***	-10.520***		-1.198	-1.009
$sMkt_{t-1}$		3.076	3.117		3.776	2.705		1.090	1.290
$fMkt_{t-1}$		-3.584	-3.493		-11.018**	-11.430**		8.828***	8.925***
$sGovt_{t-1}$		-1.585	-0.717		-0.671	-0.884		-0.046	0.064
$fGovt_{t-1}$		-6.820	-6.200		-10.377**	-10.517**		3.901	3.733
$sCorp_{t-1}$		-7.044**	-6.705**		-1.812	-0.842		-8.206**	-7.954**
$fCorp_{t-1}$		-7.703***	-7.661***		-7.280*	-6.887*		2.577	2.483
$sComms_{t-1}$		1.260	1.139		3.433	3.826		0.037	-0.661
$fComms_{t-1}$		1.802	1.732		5.054	5.240		3.193	2.766
$sMacro_{t-1}$		2.778	1.765		2.515	1.515		-1.895	-1.625
$fMacro_{t-1}$		5.865**	5.503**		3.544	3.078		2.863	3.042
$R2$	0.0697	0.127	0.125	0.0213	0.13	0.128	0.212	0.264	0.263
start	Apr 1998	May 1998	May 1998	Apr 1998	May 1998	May 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4853	4839	4839	2100	2086	2086	2753	2753	2753
$stderr$	both	both	both	both	both	both	both	both	both

Empirical findings – Panel regressions for DMs volatility

Developed markets: Forecasting panel for volatility									
	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	0.411***	0.386***	0.384***	0.354***	0.320***	0.319***	0.400***	0.372***	0.370***
$sigma_{t-2}$	0.140**	0.126**	0.126**	0.182***	0.170***	0.170***	0.075	0.044	0.044
$retmi_{t-1}$	0.664***	0.631***	0.631***	0.652***	0.605***	0.610***	0.614**	0.593**	0.592**
$retmi_{t-2}$	0.000	-0.016	-0.018	-0.074	-0.077	-0.075	0.095	0.118	0.115
$value_{t-1}$	-2.127***	-2.492***	-2.531***	-1.880*	-2.602***	-2.601***	-2.391**	-2.450**	-2.484**
gdp_{t-1}	-0.184*	-0.136	-0.129	-0.115	-0.046	-0.044	-0.176	-0.143	-0.143
$gdpdeflator_{t-1}$	-0.018	0.020	0.023	0.059	0.096	0.097	0.174	0.152	0.149
cp_{t-1}	0.010	0.006	0.006	-0.032*	-0.023	-0.022	0.030	0.033*	0.031
dcp_{t-1}	-0.018	-0.012	-0.011	0.010	0.011	0.011	-0.048**	-0.047*	-0.045*
$rate_{t-1}$	0.759***	0.835***	0.852***	1.684***	1.605***	1.615***	0.797***	0.836**	0.852**
$dexch_{t-1}$	-0.233	-0.235	-0.237	-0.524***	-0.511***	-0.516***	-0.062	-0.071	-0.072
pre	0.025	-0.006	0.003	-0.087	-0.021	-0.020	0.006	-0.179	-0.147
post	0.063	0.059	0.081	0.687*	0.843**	0.846**	-0.479	-0.531	-0.513
$entropy_{t-1}$		-2.188	-2.689		1.050	0.421		-1.974	-2.661
$artcount_{t-1}$		0.054	0.070		0.153	0.143		-0.600	-0.572
$sMkt_{t-1}$		-0.739	-0.793		-1.295*	-1.218*		-0.475	-0.616
$fMkt_{t-1}$		-0.341	-0.339		-0.348	-0.316		-0.022	-0.067
$sGovt_{t-1}$		0.616	0.565		-0.211	-0.206		1.204	1.067
$fGovt_{t-1}$		0.313	0.285		-0.165	-0.233		0.749	0.702
$sCorp_{t-1}$		0.153	0.154		0.194	0.197		-0.668	-0.614
$fCorp_{t-1}$		0.074	0.044		-0.344	-0.396		-0.416	-0.449
$sComms_{t-1}$		0.093	0.130		-0.135	-0.105		0.323	0.339
$fComms_{t-1}$		-0.240	-0.221		-0.224	-0.230		-0.295	-0.275
$sCredit_{t-1}$		-0.604	-0.647		0.087	-0.029		-0.402	-0.373
$fCredit_{t-1}$		-0.087	-0.107		0.334	0.309		-0.417	-0.424
R2	0.473	0.479	0.479	0.466	0.475	0.475	0.453	0.459	0.459
start	Apr 1998	May 1998	May 1998	Apr 1998	May 1998	May 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4422	4406	4406	2003	1987	1987	2419	2419	2419
stderr	by time	by time	by time	by time	by time	by time	by time	by time	by time

Empirical findings – Panel regressions for DMs drawdowns.

Developed markets: Forecasting panel for drawdowns

	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	0.098	0.016	0.011	0.068	0.033	0.035	-0.005	-0.099	-0.094
$sigma_{t-2}$	0.171**	0.110	0.112	0.156**	0.097	0.095	0.081	-0.014	-0.003
$return_{t-1}$	-0.392***	-0.261**	-0.264**	-0.186*	-0.091	-0.095	-0.414***	-0.268*	-0.272*
$return_{t-2}$	-0.073	-0.008	-0.007	-0.031	-0.011	-0.010	-0.065	-0.016	-0.012
$value_{t-1}$	-12.271***	-13.805***	-13.878***	-6.073***	-6.611***	-6.589***	-15.672***	-16.588***	-16.749***
gdp_{t-1}	0.165	0.412	0.445	-0.092	0.040	0.071	0.792**	0.794**	0.780**
$gdpdeflator_{t-1}$	-0.286	-0.081	-0.069	-0.349	-0.241	-0.218	0.361	0.156	0.149
cp_{t-1}	0.185***	0.153***	0.156***	-0.008	0.033	0.027	0.149**	0.144**	0.144**
dcp_{t-1}	-0.070	-0.038	-0.034	0.074	0.068	0.068	-0.144	-0.115	-0.113
$rate_{t-1}$	3.057***	3.885***	3.964***	7.657***	7.476***	7.403***	4.185***	4.427***	4.535***
$dexch_{t-1}$	-0.398	-0.334	-0.343	-0.094	-0.001	-0.008	-0.471	-0.424	-0.430
pre	1.376	1.947	1.981	2.030	2.089	2.096	0.065	0.250	0.240
post	0.128	0.351	0.391	0.484	0.848	0.822	-0.572	-1.020	-1.060
$entropy_{t-1}$		-11.165*	-13.357**		11.666*	9.826		-12.148*	-13.010**
$artcount_{t-1}$		0.149	0.217		-0.057	-0.001		-0.840	-0.819
$sMkt_{t-1}$		-4.132***	-4.348***		-0.499	0.080		-6.765***	-7.292***
$fMkt_{t-1}$		-0.866	-0.893		0.589	0.715		-1.330	-1.388
$sGovt_{t-1}$		4.576***	4.303***		0.551	0.708		6.013***	4.945***
$fGovt_{t-1}$		3.860***	3.818***		0.857	0.938		5.574***	5.261***
$sCorp_{t-1}$		1.195	0.632		-1.537*	-2.011**		0.343	0.484
$fCorp_{t-1}$		2.614**	2.323**		1.381	1.267		-0.062	-0.310
$sComms_{t-1}$		-0.783	-0.509		-0.654	-0.681		0.281	0.410
$fComms_{t-1}$		-0.098	-0.022		0.386	0.291		0.340	0.416
$sCredit_{t-1}$		-1.654	-1.207		-0.252	-0.404		1.610*	2.341***
$fCredit_{t-1}$		0.382	0.379		-0.429	-0.424		0.868	0.944
R2	0.263	0.323	0.322	0.404	0.45	0.453	0.389	0.448	0.448
start	Apr 1998	May 1998	May 1998	Apr 1998	May 1998	May 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4422	4406	4406	2003	1987	1987	2419	2419	2419
stderr	both	both	both	both	both	both	both	both	both

Empirical findings – Summary

- We detect the connections between various measures of word flow and our measures of expected return, the standard deviation of returns (sigma), and cumulative downside risk (drawdown).
- *Return , sigma,* and *drawdown* tend to be more predictable for DMs. → The nature of news, and the range of potential news outcomes, differ in EMs and DMs.
- When a word flow measure has a positive (negative) effect on return , it often tends to have a negative (positive) effect on sigma and a negative effect on drawdown. → The factors captured by news flow are not priced risks.
- The incremental contribution to R-squared of word flow measures tends to be relatively small for return and sigma, compared to that to drawdown. → The nature of news tends to be different in EMs and DMs

Empirical findings – Summary

- Effects of specific text measures. → The impacts of individual text flow measures on annual returns and drawdowns often are economically large.
- Entropy interactions. → We do not find that interacting sentiment measures with entropy, the *SentEnt* specification, adds much explanatory power.
- Time variation in coefficients. → Consistent with our principal component discussion, we find important differences in coefficient values for word flow measures over time.
- Sign of sentiment and market outcomes. → Coefficients for sentiment or frequency can be positive or negative. There is no general finding that positive sentiment is always associated with good news.

Empirical findings – Pre- and postcrisis differences in the meaning of news flow

- We investigate whether post-crisis differences reflect changes that persist throughout the period or changes that are only related to the onset of the global financial crisis.

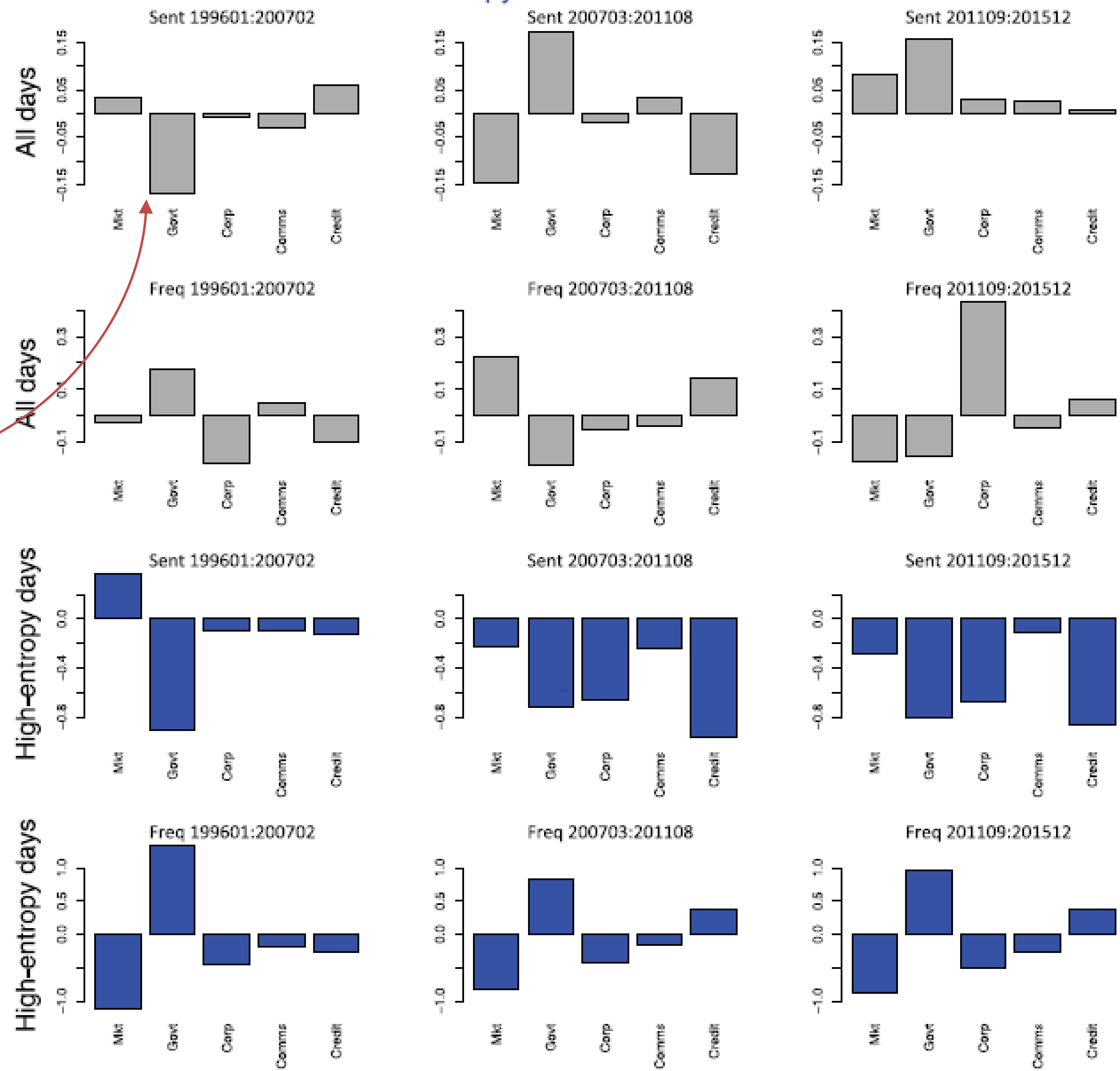
- To examine the nature of the role of crisis influences, we divide the post-February 2007 time period into two subperiods: the global crisis period from March 2007 to August 2011 (the midpoint of the post-February 2007 period) and the subperiod after August 2011.

Entropy statistics for DM

Each chart shows the difference between the average country-day sentiment or frequency in that subperiod/ entropy grouping and the full-sample average.

average government sentiment was 0.15 standard deviations lower than the full-sample average

→ It appears that the changes in the structure and content of news related to the onset of the crisis were more persistent in DMs, where the crisis and policy reactions to it were more long lasting.



Out-of-sample tests

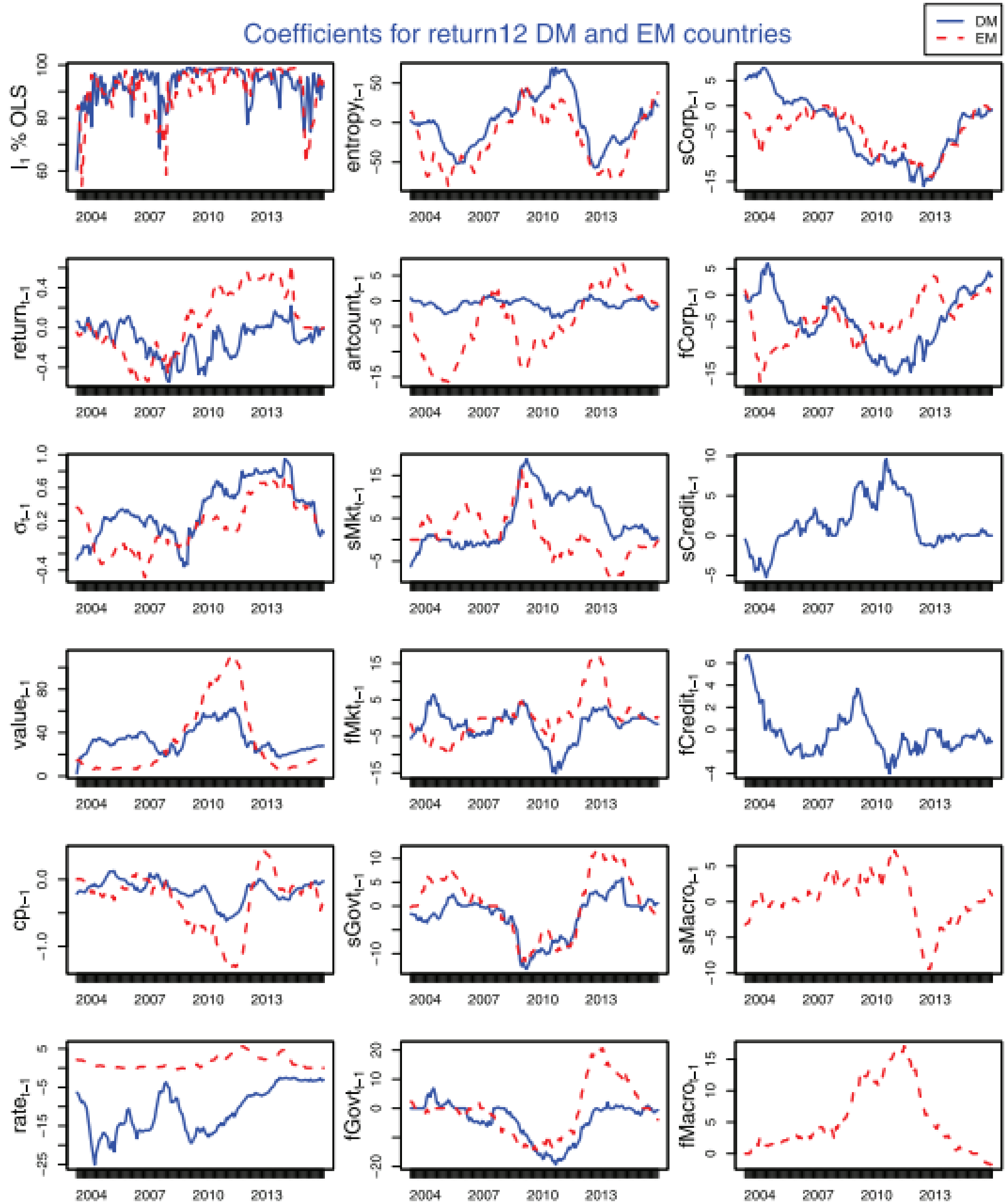
① the substantial variation over time in coefficient estimates reported

② the baseline and augmented models contain many explanatory text and non-text variables that make them susceptible to overfitting in any given sample → the elastic net estimator, which combines the least absolute shrinkage and selection operator (lasso) regression with a ridge regression

③ confronted with too many explanatory variables and a relatively small data set → use only a subset of our nontext variables for the out-of-sample tests

Out-of-sample tests - the coefficient estimates from a rolling elastic net regression to forecast 12-month returns

- Positive sMkt tends to be good news for future market outcomes, whereas positive sGovt, fGovt, sCorp and fCorp tend to be bad news.



Trading strategy based on out-of-sample model predictions

- We analyze how useful textual information would be to a mean-variance optimizing investor who already had access to our baseline model's out-of-sample forecasts.
- The three forecasting models are:

① Naive, which uses only in sample country fixed effects as the forecasting variables;

② Base, which includes lagged macroeconomic and lagged market variables as the regressors;

③ CM, which includes country specific article counts, entropy, sentiment, and frequency measures in addition to the variables from the Base model.

Trading strategy based on out-of-sample model predictions

Following Campbell and Thompson (CT, 2008), we assume a myopic mean-variance investor.

Both differences are clearly important economically, especially so for EM countries, indicating the improved investment performance.

Panel A

Developed market strategy

Model	Alpha	Mkt.RF	fxcarry	fxusd
CM	8.816	0.443	-0.168	-0.413
	(1.438)	(2.702)	(-0.591)	(-2.756)
Base	6.809	0.570	-0.076	-0.395
	(1.380)	(4.286)	(-0.347)	(-2.784)
Naive	-2.765	0.666	0.123	-0.024
	(-0.678)	(4.635)	(0.749)	(-0.219)

Emerging market strategy

Model	Alpha	Mkt.RF	fxcarry	fxusd
CM	8.801	0.358	0.103	-0.298
	(1.960)	(2.621)	(0.591)	(-2.235)
Base	3.271	0.499	0.137	-0.318
	(0.892)	(4.677)	(1.158)	(-2.645)
Naive	2.347	0.529	0.334	-0.175
	(0.780)	(5.370)	(2.281)	(-1.766)

Panel B

Tests comparing alphas of CM and Base models

Market	Difference in alphas/yr	*T*-test *p*-values	
		2-sided	1-sided
DM	2.01	0.082	0.041
EM	5.53	0.002	0.001

Conclusion

- We develop an atheoretical approach for capturing news through various word flow measures and apply that approach to 51 countries over the time period 1998–2015.
- We find that news contained in our text flow measures forecasts one-year ahead returns and drawdowns.
- Basic statistical properties of news and returns are different for EMs and DMs, as are the relevant topics for news stories.
- We find that coefficient values on various word flow measures do change over time.
- We perform out-of-sample testing using an elastic net regression to investigate whether our model is economically useful.