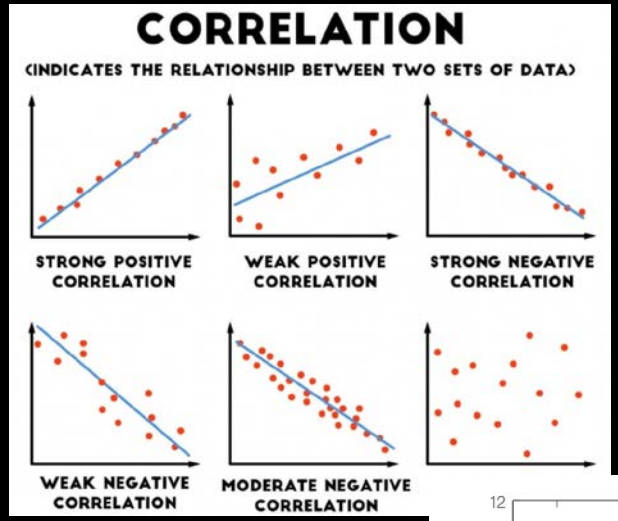
Predictability Using Big Data

Discussion by
Sanjiv R Das
Santa Clara University

Levels of Dependence



Prediction

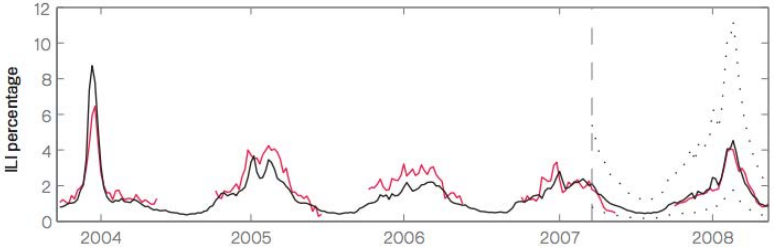
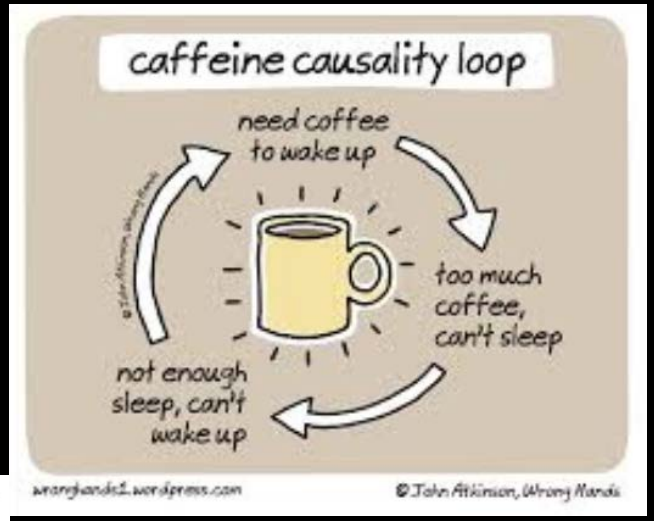


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.



Causality

Sentiment Scoring

Loughran and McDonald JF 2011 word lists

- Fin-Neg – negative words (e.g., *loss, bankruptcy, indebtedness, felony, misstated, discontinued, expire, unable)*. N=2,349
- Fin-Pos – positive words (e.g., *beneficial*, *excellent*, *innovative*). N = 354
- Fin-Unc – uncertainty words. Note here the emphasis is more so on uncertainty than risk (e.g., ambiguity, approximate, assume, risk). N =291
- Fin-Lit – litigious words (e.g., admission, breach, defendant, plaintiff, remand, testimony). N =871

Negation and Stemming

- *Negation tagging*: Notice that in financial reporting it is unlikely that negative words will be negated (e.g., not terrible earnings), whereas positive words are easily qualified or compromised. Although you can easily account for simple negation, typical forms of negation are difficult to detect.
- *Stemming* does not work well for morphologically rich languages
- *Relevance* and *Novelty*: do these features matter? ("unusualness")

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN (CM)	MEDIA REINFORCEMENT EFFECTS (FLOSS)

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN (CM)	MEDIA REINFORCEMENT EFFECTS (FLOSS)
Predictive variables	Topic-specific sentiment, frequency, unusualness of word flow	Media measured sentiment, change compared to past 4 weeks

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN (CM)	MEDIA REINFORCEMENT EFFECTS (FLOSS)
Predictive variables	Topic-specific sentiment, frequency, unusualness of word flow	Media measured sentiment, change compared to past 4 weeks
Predicted variables	Country-level returns, volatilities, drawdowns	Returns (equities and currencies)

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN (CM)	MEDIA REINFORCEMENT EFFECTS (FLOSS)
Predictive variables	Topic-specific sentiment, frequency, unusualness of word flow	Media measured sentiment, change compared to past 4 weeks
Predicted variables	Country-level returns, volatilities, drawdowns	Returns (equities and currencies)
Time scale	• Monthly, annual • Stationarity?	Weekly

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN (CM)	MEDIA REINFORCEMENT EFFECTS (FLOSS)
Predictive variables	Topic-specific sentiment, frequency, unusualness of word flow	Media measured sentiment, change compared to past 4 weeks
Predicted variables	Country-level returns, volatilities, drawdowns	Returns (equities and currencies)
Time scale	• Monthly, annual • Stationarity?	Weekly
Stronger effect	Emerging markets (less liquid)	For liquid markets

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN	MEDIA REINFORCEMENT EFFECTS
Sample size	52 countries, developed and emerging markets (non-US)	Developed country markets, mostly liquid

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN	MEDIA REINFORCEMENT EFFECTS
Sample size	52 countries, developed and emerging markets (non-US)	Developed country markets, mostly liquid
News source	Thomson-Reuters from 1996 to 2015	Cross-asset media coverage (01/2013 - 03/2017); multiple news sources

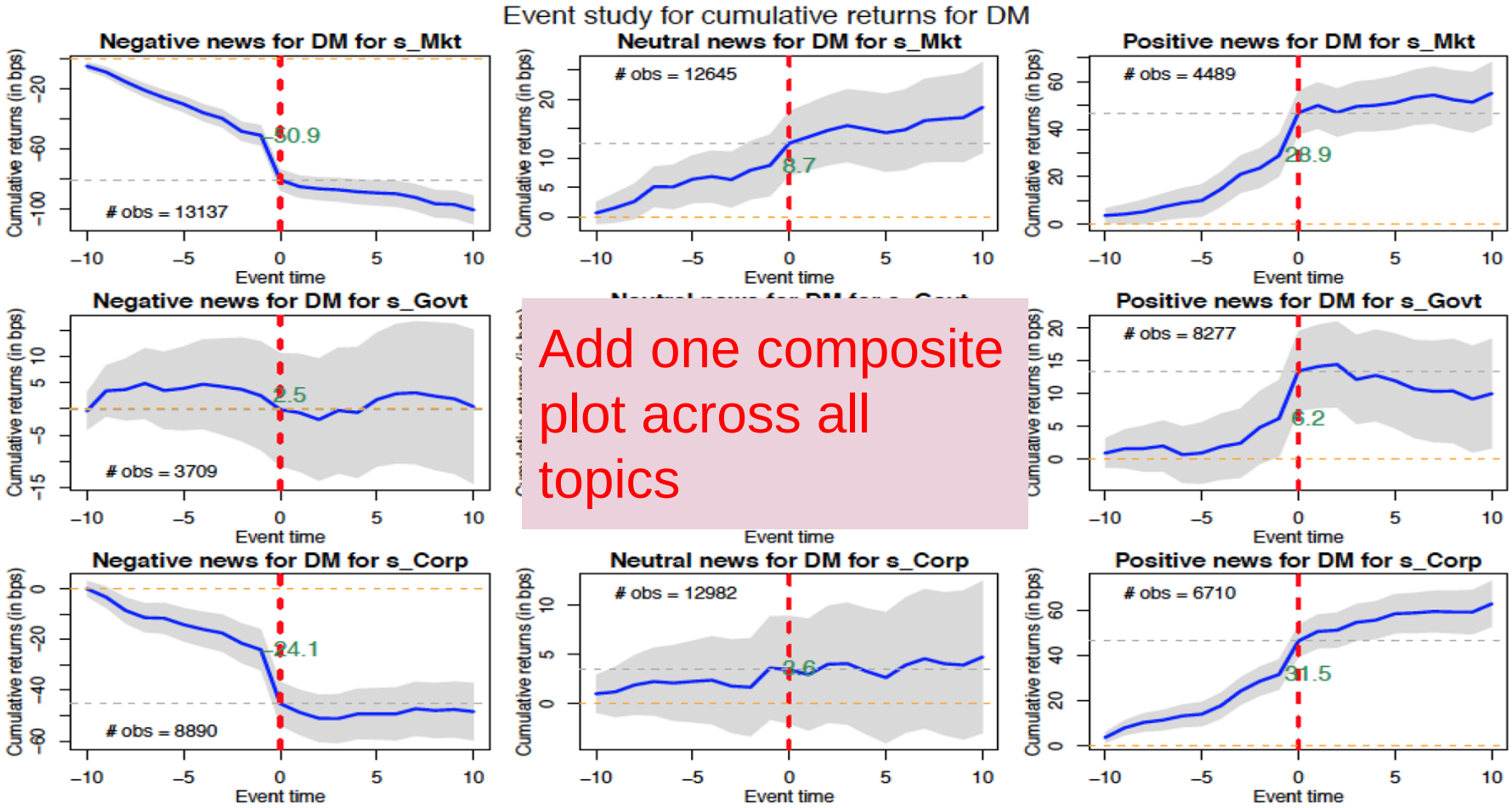
Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN	MEDIA REINFORCEMENT EFFECTS
Sample size	52 countries, developed and emerging markets (non-US)	Developed country markets, mostly liquid
News source	Thomson-Reuters from 1996 to 2015	Cross-asset media coverage (01/2013 - 03/2017); multiple news sources
Sentiment	● Senty = (Pos - Neg)/(Total) ● Interact with topic ● Entropy	● Senty = (Pos - Neg)/(Pos + Neg) ● Interact senty with direction

Comparison of the two papers

ITEM	HOW NEWS DRIVES RISK AND RETURN	MEDIA REINFORCEMENT EFFECTS
Sample size	52 countries, developed and emerging markets (non-US)	Developed country markets, mostly liquid
News source	Thomson-Reuters from 1996 to 2015	Cross-asset media coverage (01/2013 - 03/2017); multiple news sources
Sentiment	• Senty = (Pos - Neg)/(Total) • Interact with topic • Entropy	• Senty = (Pos - Neg)/(Pos + Neg) • Interact senty with direction
Methodology	• Event study, cumulative abnormal returns • Regression analysis • Lasso (overfitting)	• Calendar time method • Regressions to get alpha

Event Study Analysis



Add one composite plot across all topics

Suggestions : CM paper

- Show samples of words that make up sentiment most strongly
- Reverse causality? Does the market affect media?
- Non-stationarity? (with text this may be less of a concern)
- Both papers focus on returns (and risk) but not on major event prediction, for which see the Enron analysis.
- Look both ways: (i) find big sentiment swings and see if they are followed by large return moves (done); (ii) find big return moves and see how many are preceded by sentiment swings. (These two things are different.) CM paper does have an event study focus, but if the annual effect is stronger, why not show this also?
- Possible to implement a trading strategy (as in the FLOSS paper)
- Signal "decay" - how does the IR of the strategy drop as the aggregation window increases?
- t-SNE for graphical depictions of different types of textual regimes.

t-SNE

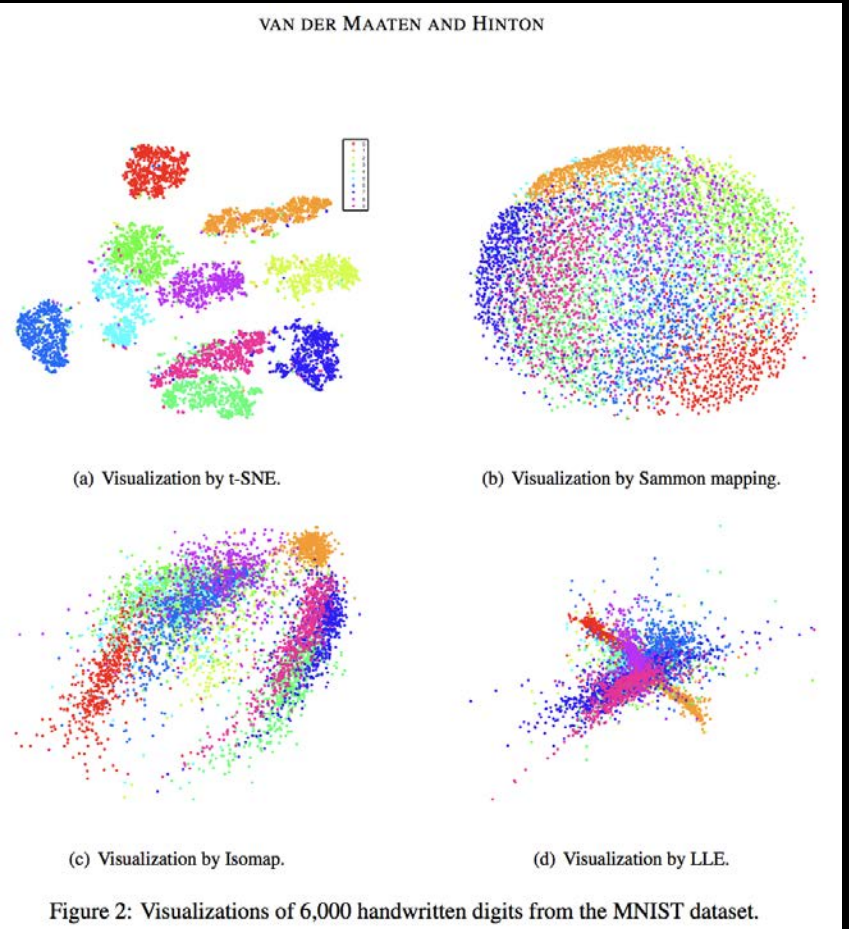
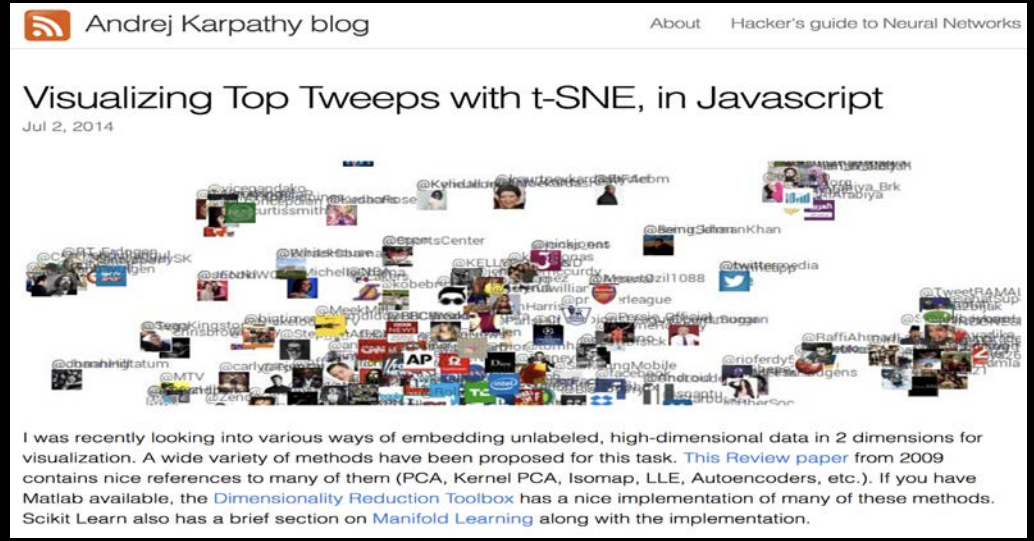


Figure 2: Visualizations of 6,000 handwritten digits from the MNIST dataset.

- t-SNE (van der Maaten and Hinton, 2008) is based on Stochastic Neighbor Embedding developed by Hinton and Rouweis (2002)
- Visualizes high-dimensional data on a two or three dimensional map
- Use word2vec to create word vectors



http://cs.stanford.edu/people/karpathy/tsnejs/

Suggestions : FLOSS paper

- How long does each condition persist in the data set? Does it affect the weekly regressions, and do we need lags?
- 4 types : returns hi +low; media hi +low. Show these with colored paths. What is the transition matrix for these?
- Effects on volatility (risk)?
- Effect on media sentiment from returns? (VAR)
- Randomized experiment with leads and lags? Can we improve on this result?
- Look at return distributions to examine higher order moments.
- Is there an option timing argument that can be made here? This really is market timing with a specific signal mechanism. How much is the timing option worth?
- Has the reinforcement effect become stronger over time? Is the media effect growing (fake news)?
- Report turnover rates for the portfolio strategy.
- Benchmarks? (for IR)

Industry Practice : Ravenpack

Sources: Dow-Jones, WSJ, Barron's, and 19000 other traditional and social media sites.

USFAST Model	Large/Mid-Cap		Small-Cap	
Statistics	Specific Ret	Excess Ret	Specific Ret	Excess Ret
Annualized Return	12.89%	10.89%	32.66%	30.13%
Annualized Vol.	3.43%	4.50%	5.38%	6.12%
Information Ratio	3.76	2.42	6.07	4.92
Avg. Portfolio Size	188	190	193	193
Max. Drawdown	2.31%	6.40%	2.81%	4.42%
Turnover	85.2%	83%	89.9%	89.5%

TABLE 1: Experimental Stats. Comparison between MSCI specific returns and excess returns for a set of statistics. Results are shown for USFAST model for Large/Mid-Cap and Small-Cap.

Source: RavenPack, MSCI, October 2017

https://www.ravenpack.com/files/research/sentiment-signals-msci-barra-risk-models/?utm_campaign=msci&utm_medium=email&utm_source=link&utm_content=&utm_term=

Performance

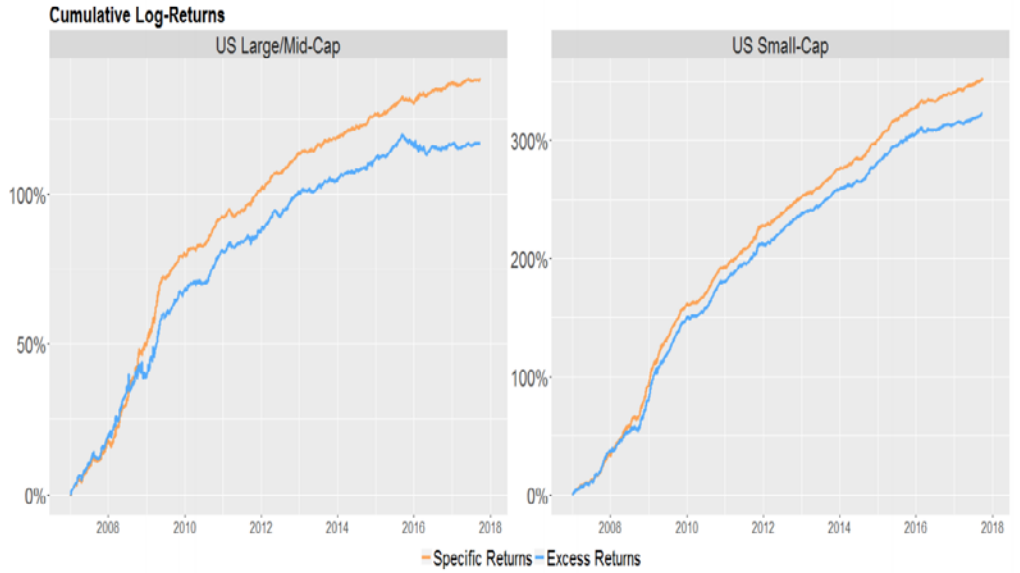


FIGURE 4: Cumulative Return Profile Comparison for USFAST Model. Comparison between MSCI specific returns and excess returns for a daily strategy on a one-day aggregated RPA signal. Results are shown U.S. Large/Mid-Cap (left) and Small-Cap (right) Universes.

Source: RavenPack, MSCI, October 2017

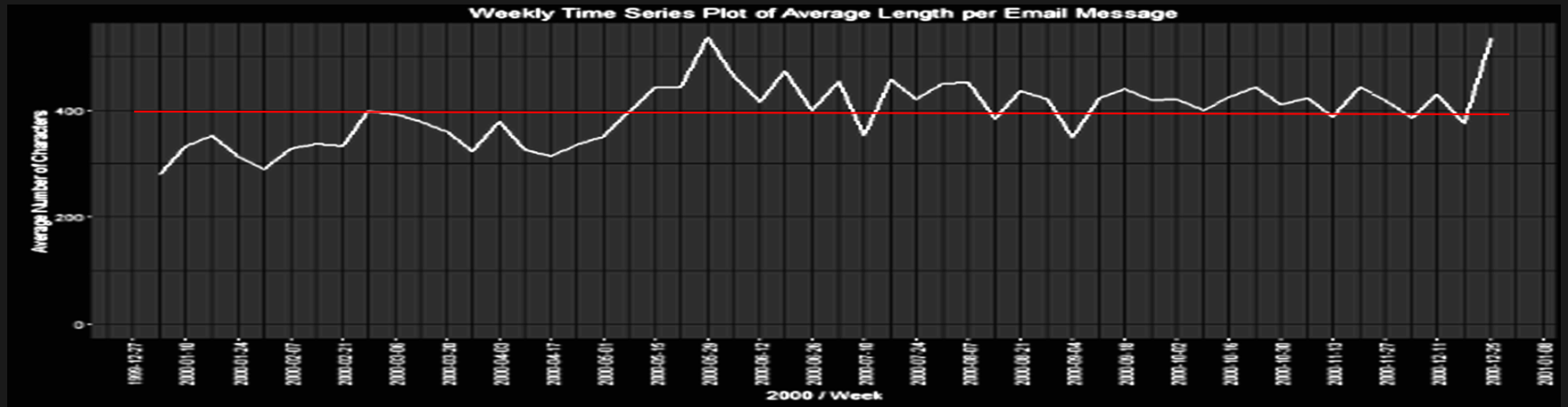
- For the U.S., Information Ratios (IR) of 3.8 and 6.1 are achieved for the Russell 1000 and the Russell 2000, respectively. We get IRs of 3.2 and 4.2 for their European equivalent.
- The predictive power of the signal is statistically significant and positive across the entire backtesting period.
- The prediction quantile analysis shows a desirable profile, providing higher returns for more extreme predictions.

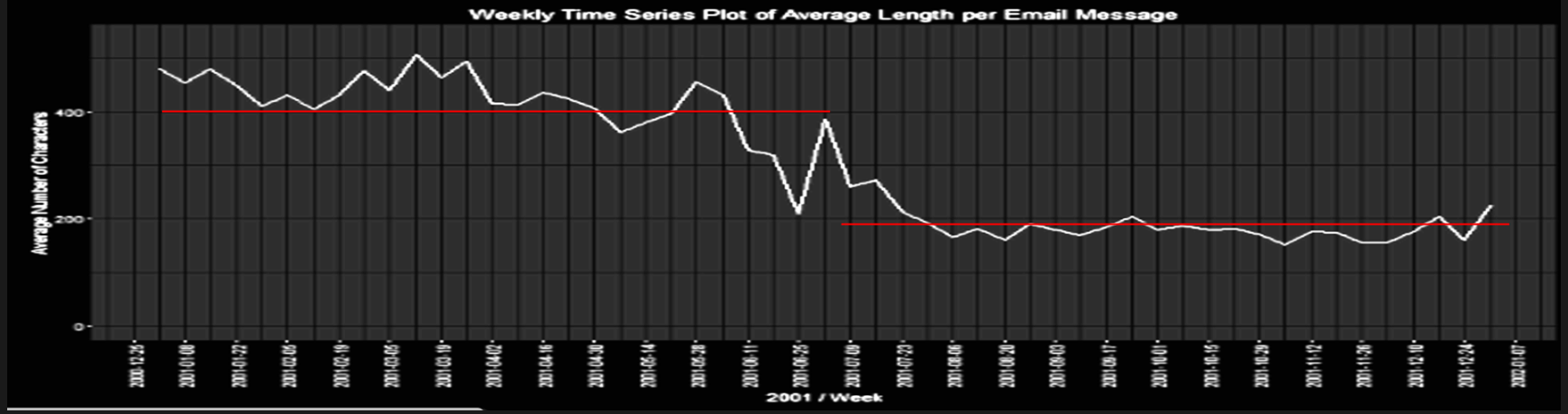
RegTech

Zero-Revelation Linguistic Regulation: Detecting Risk Through Corporate Emails and News (Das, Kim, Kothari 2016)

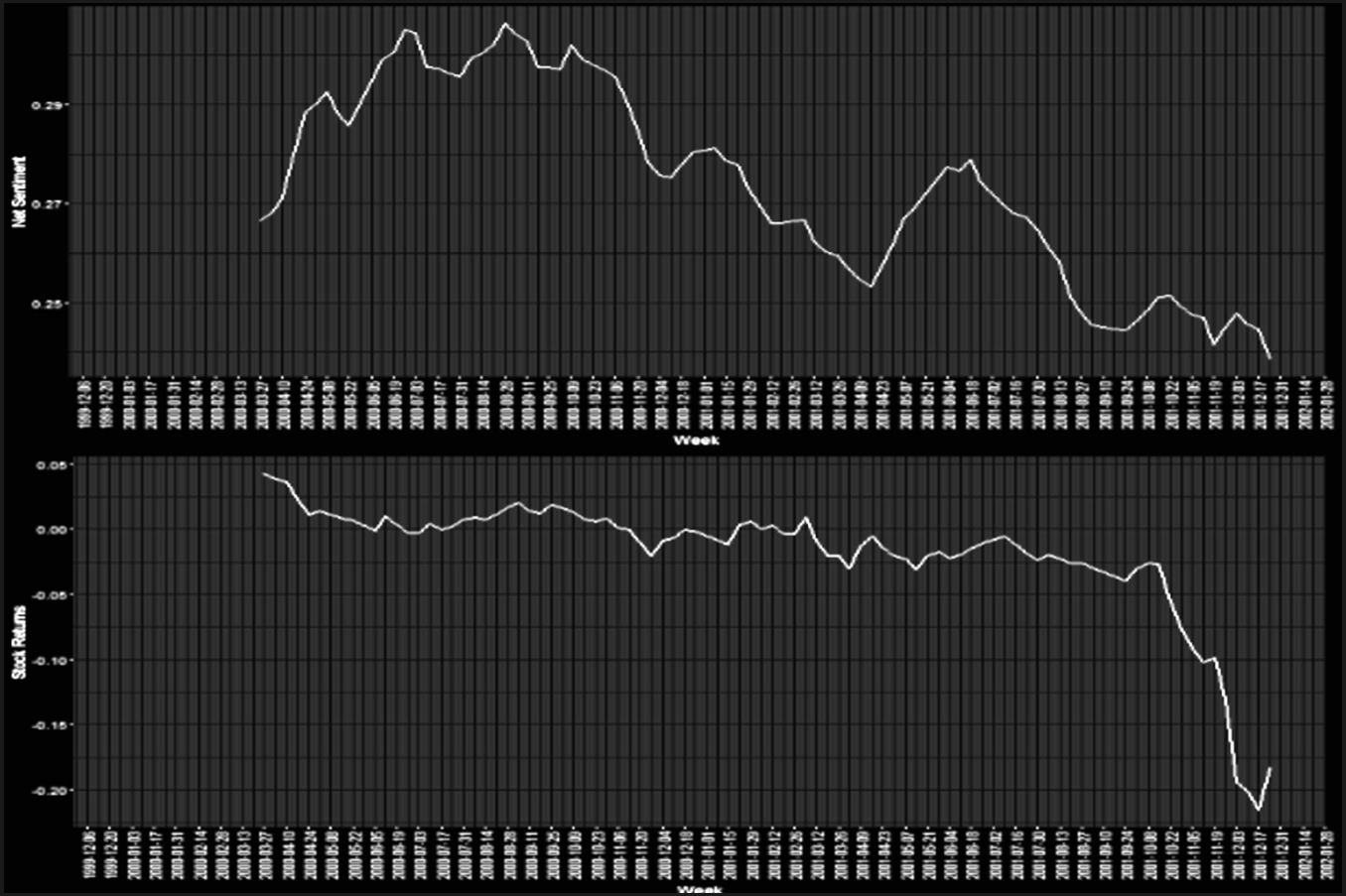
- Financials are often delayed indicators of corporate quality.
- Internal discussion may be used as an early warning system for upcoming corporate malaise.
- Emails have the potential to predict such events.
- Software can analyze vast quantities of textual data not amenable to human processing.
- Corporate senior management may also use these analyses to better predict and manage impending crisis for their firms.
- The approach requires zero revelation of emails.

Enron: Email Length



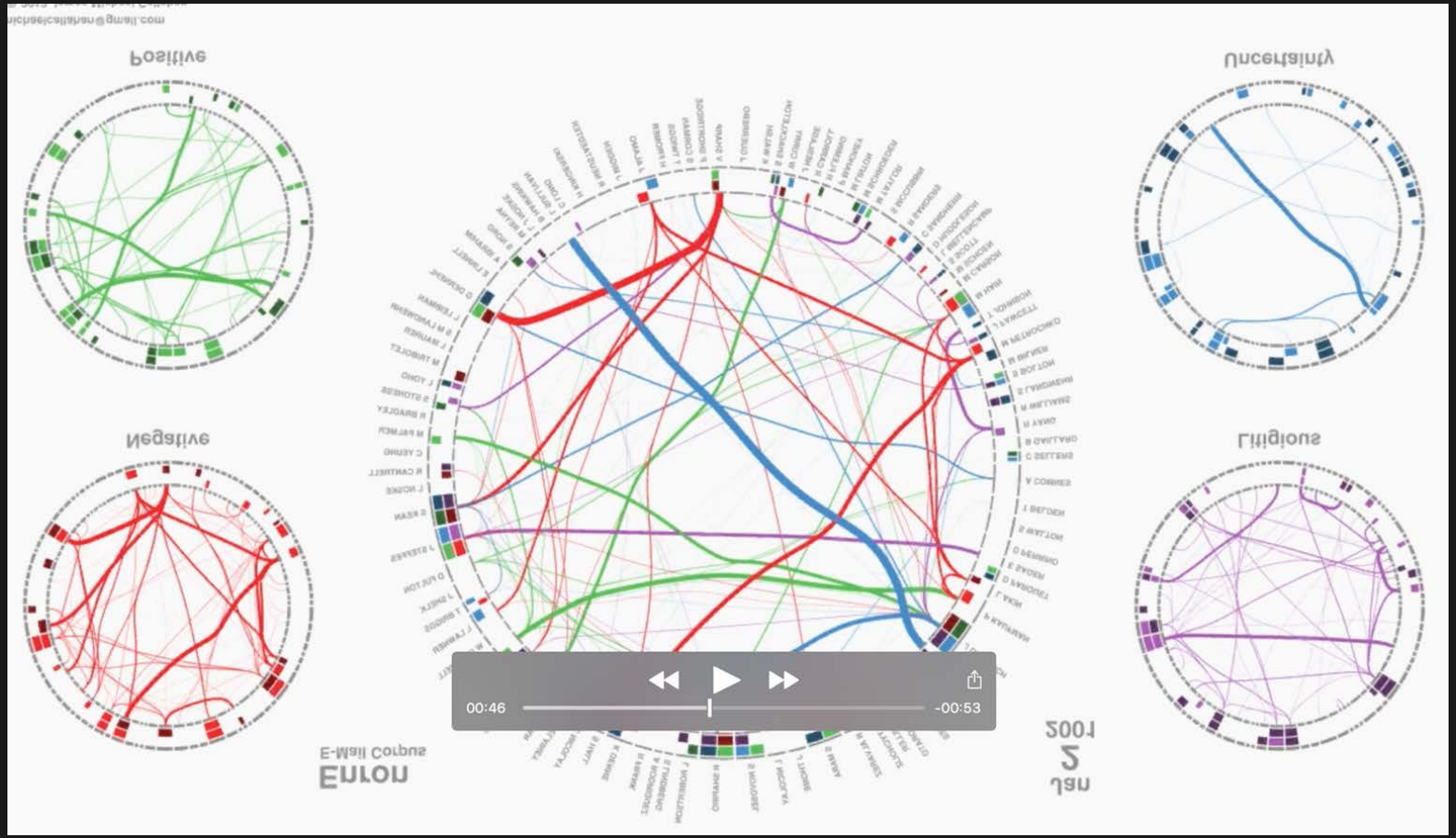


Enron: Sentiment and Returns



Enron: Returns and Characteristics

Variable	Coefficient Estimate (*t*-statistic)			
	(1)	(2)	(3)	(4)
MA Net Sentiment $_t$	XXX***	0.575	2.330***	-1.397
	(XXX)	(0.63)	(3.14)	(-1.25)
MA Email Length $_t$		0.584***		1.046***
		(2.97)		(4.19)
MA Total Emails $_t$			-0.004	-0.131***
			(-0.10)	(-2.83)
Intercept		-0.406*	-0.671***	0.117
		(-1.93)	(-3.08)	(0.43)
Adjusted *R*-squared	XXX		0.09	0.24
Number of observations	88	88	88	88

Enron Movie (by Jim Callahan)

http://srdas.github.io/Presentations/JimCallahan_enron-sm.mov