The Stock Market Impact of the Foreign and the Domestic Media

by

Gábor Nyéki

Submitted to

Central European University

Department of Economics

In partial fulfilment of the requirements for the degree of

MA in Economics

Supervisor: Ádám Szeidl

Budapest, Hungary

2012

I, the undersigned [Gábor Nyéki], candidate for the degree of MA in Economics at the Central European University Department of Economics, declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of works of others, and no part the thesis infringes on any person's or intstitution's copyright. I also declare that no part the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Budapest, 4 June 2012

Signature

Acknowledgements

I thank Ádám Szeidl and László Mátyás for their suggestions about this thesis.

I also appreciate the comments of my classmates in our thesis discussions.

Table of Contents

Chapter 1

Introduction

Do foreign news matter for domestic investments? The answer's relevant for the theory of market efficiency as well as for the practice of communication in business and in politics. This thesis is an investigation of the issue. I study stock price movements of Indian companies, relating them to mentions in the Indian Economic Times and the British Financial Times.

I test three hypotheses: (1) efficient markets, (2) media impact on stock markets, and (3) foreign media impact. Stock market data shows patterns that suggest both general media impact and foreign media impact in particular. Price behaviour around company mentions also provides evidence against efficient markets.

I distinguish between the unobserved underlying *events* and *media mentions.* To spot mentions, I collect news from the online edition of the two newspapers, and company data from the stock quote catalogue of the Economic Times. For most of these tasks, I use an open-source web crawling framework for the Python scripting language. The size of these websites and the time available mean that only a fraction of the news that were published are reached. Over four weeks of news collection, I found approximately 100,000 articles overall. In these articles, I identify 4,692 Indian and 369 British mentions from 2007 to 2012, involving 462 and 57 companies, respectively. These mentions are matched with stock market data from Yahoo! Finance and the Bombay Stock Exchange.

It is evident that there is a relationship between Indian mentions and price movements. Looking at single Economic Times mentions, absolute market-adjusted returns are higher than usual one day before the mention and on the day of the mention. The relationship between mentions and trading activity is further illustrated by simple correlation coefficients between market-adjusted returns one day before and on mention days. For Indian mentions, the coefficient is .049 (p-value = .01). For British mentions, it is .0902 (p-value = .08) for all mentions, and $r = .2206$ (p-value = .01) for only mentions that are not preceded by Indian mentions.

The opposite pattern is observed with subsequent days with no mentions. For those, the coefficient is $-.037$ (p-value $< .001$). This daily return reversal is in contrast with the price momentum in the presence of mentions.

These patterns are robust. I also capture them by regressing market-adjusted returns on mentions and lagged returns. The results from the correlation coefficients are preserved in the regressions, with and without controls as well. Overall the evidence leaves space for speculation about foreign investor behaviour. What is a fact is that Financial Times mentions are associated with daily return reversal—the same behaviour as without any mentions—if they follow Indian mentions.

Due to missing data, my results are affected by attenuation bias. The bias could be lessened by restricting the sample to periods with more observed mentions. However, this also severely reduces the sample size. Noting these, while the Financial Times appears to have more noise in mentions, the basic results are still robust across different tests.

1.1 Literature

The effect of financial news on stock markets has been studied by several authors. The closest to this thesis are Tetlock (2011) and Shabani (2011). Tetlock looks at market response to “stale news”—news containing information published earlier. He measures staleness by textual similarity, and finds response as well as a subsequent return reversal. While Tetlock uses daily price data, Shabani generates a timestamped transcript of CNBC’s television broadcast and works with the minute-by-minute breakdown of price movements. He restricts his attention to earnings announcements, defining them as events, and compares market response to events with response to CNBC mentions. Shabani also finds trade on stale information, although no evidence of subsequent return reversal.

Beyond these papers, Engelberg and Parsons (2011) look at response to local newspaper coverage of earnings announcements. They find a relationship between media coverage and trading activity. This result is supported by an exercise introducing extreme weather conditions as exogenous variation in coverage. Importantly, their sample spans from 1991 to 2007, so much of it is from when the internet didn’t yet reach universality. Finally, Dougal et al. (2012) examine the individual impact of Wall Street Journal columnists on market activity. They find a strong relationship, telling apart bullish and bearish journalists based on their impact. Their sample period spans from 1970 to 2007, so their data is also mostly from before the internet era.

I extend on these by considering foreign as well as domestic mentions. My approach explicitly relies on the prevalence of online media as I ignore print publication dates. Note, however, that the observed practice of the Financial

Times, as well as for instance of the Australian Financial Review which is not studied deeper in this thesis, is to publish news appearing in the print edition on their online outlet the night before. Indeed, news competition also dictates this behaviour.

Chapter 2

Data

I use two kinds of data for the analysis, financial news and stock prices. I collect the news from the Economic Times (India) and the Financial Times (United Kingdom). For data on stock prices, I track the Bombay Stock Exchange and companies listed there. Data on prices is daily.

Regarding terminology, throughout the thesis I use *mention* to refer to news mentioning the name of a certain company, and *event* to refer to the underlying event generating the mentions. Therefore mentions are observable and events are not. Figure 2.2 illustrates how events are often likely to be not accompanied by mentions in the sample. This is due to the enormous amount of news published by the Economic Times and the Financial Times, and the limitations of the data collecting process.

2.1 Companies and stock prices

I am only looking at mentions of companies listed on the Bombay Stock Exchange (BSE). Company names and matching ticker codes were obtained from the website of the Economic Times. This way I collected data on roughly 2,700 companies.

I downloaded daily price data on individual stocks from Yahoo! Finance. Yahoo! could not find price data for about a third of the ticker codes, without any noticeable pattern in the omissions. Due to this, I could match only 1,866 BSE tickers with price data.

Beside stock prices, I also collected daily data on the BSE 500 market index. It is running from 2005 to ensure it is not a constraint when searching the news for company mentions.

In this thesis, the main variable of interest is abnormal (i.e., market-adjusted) returns. This is to obtain a normalised measure of price changes. Abnormal returns are defined as

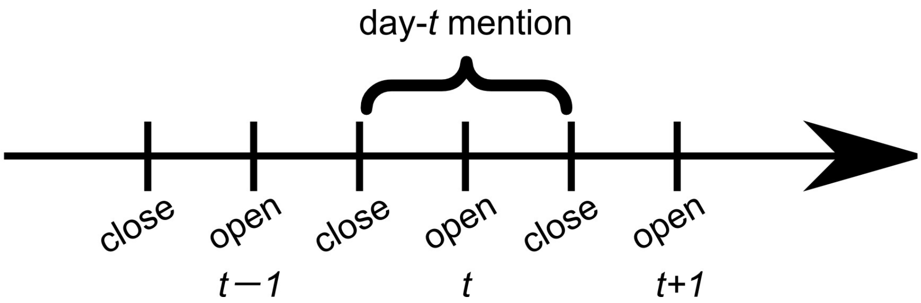
$$AR_{it} \equiv \ln\left(\frac{\text{price}_{it}}{\text{price}_{it-1}}\right) - \ln\left(\frac{\text{BSE500}_t}{\text{BSE500}_{t-1}}\right),$$

So values of $100 \times AR_{it}$ around zero are approximately equal to the deviation from market returns, in percentages. Absolute abnormal returns is a measure of price volatility and reflects trading activity. In this study, I am looking at closing prices.

2.2 News

The timing of mentions is depicted in Figure 2.1. If they were published before the stock exchange closed on trading day t, they are linked to that trading day. This means that news that correspond to day t came out in the 24 hours after

Figure 2.1: Timing of company mentions



Note: Mentions are classified as having happened on day t if they occurred after the stock market closing on day $t-1$ and before closing on day t. Accordingly, abnormal returns are computed using closing prices.

Table 2.1: Mentions and companies in the sample

Panel A: Number of mentions

	all	Jan, 2011–Apr, 2012	Jan–Jun, 2011
Economic Times (India)	4,692	3,301	1,523
Financial Times (UK)	369	129	56

Panel B: Number of companies mentioned

	all	Jan, 2011–Apr, 2012	Jan–Jun, 2011
Economic Times (India)	462	371	269
Financial Times (UK)	57	32	22

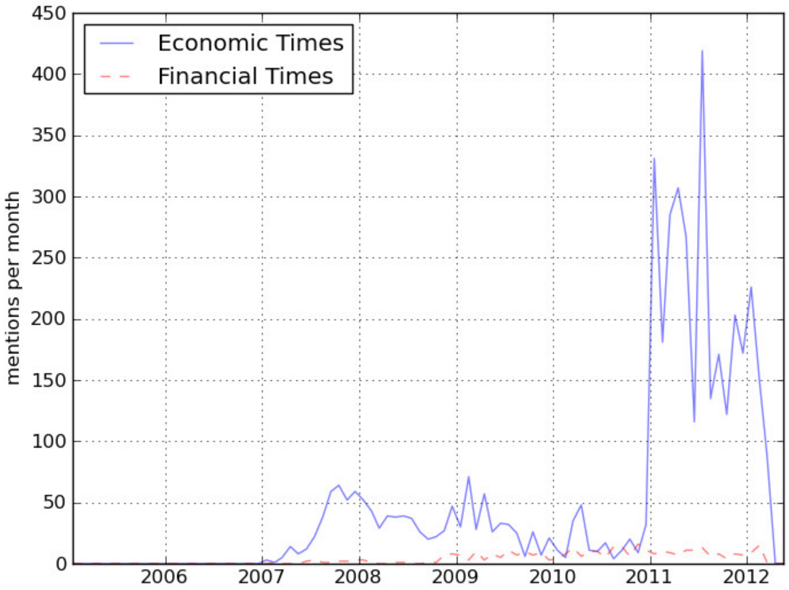
Note: The problem of measurement error (unobserved mentions) could be alleviated by restricting the sample at the heavy cost of losing observations.

the market closed on day $t-1$.

The news collecting process, to save resources by not purchasing the news, was essentially what is called “crawling” the websites of these newspapers. A web crawler is the eyes and ears of every search engine—starting off somewhere on the internet, it follows links and processes pages it finds on the way. Using the open-source Scrapy framework for Python, I programmed a crawler for the Economic Times and the Financial Times to dig up company mentions.

The choice of the British Financial Times and the Indian Economic Times is expected to be representative of the financial media in these countries. The

Figure 2.2: Distribution of mentions over time in the sample



Note: The figure shows mentions each month in the sample. Many of the company mentions that occurred are likely missing. This measurement error may cause attenuation bias.

Economic Times is the largest such newspaper in India, with a reader base of about 800 thousand people. Both the Financial Times and the Economic Times publish at least an excerpt of their content on their websites. My point of reference is the time of online publication—I don't consider the print edition.

The drawback of web crawling for finding company mentions is that if the tedious crawling procedure is not finished, that is, if not every link is visited, entire regions of the website can be left unexplored. What this means in the case of news is illustrated in Figure 2.2.

Because the online editions of the Financial Times and the Economic Times are a lot more extensive than what can be realistically scraped on a portable computer, the analysis has got to be carried out on incomplete data. This introduces two kinds of problems, one is measurement error (we believe there was no mention of a company on day t whereas there was), another is Arthur Goldberger's micronumerosity—or the relatively small size of the sample.

Overall, I have collected 54,715 news items from the Financial Times and 44,779 news items from the Economic Times. Among these, 553 in the Financial Times and 14,052 in the Economic Times mentioned companies that could be matched with ticker codes on the Bombay Stock Exchange. This number is further reduced because of Yahoo!'s failure to find some of the companies, and because several of these mentions occurred during what is classified as the same trading day. The final number of mentions is detailed in Table 2.1.

2.3 Patterns in trading

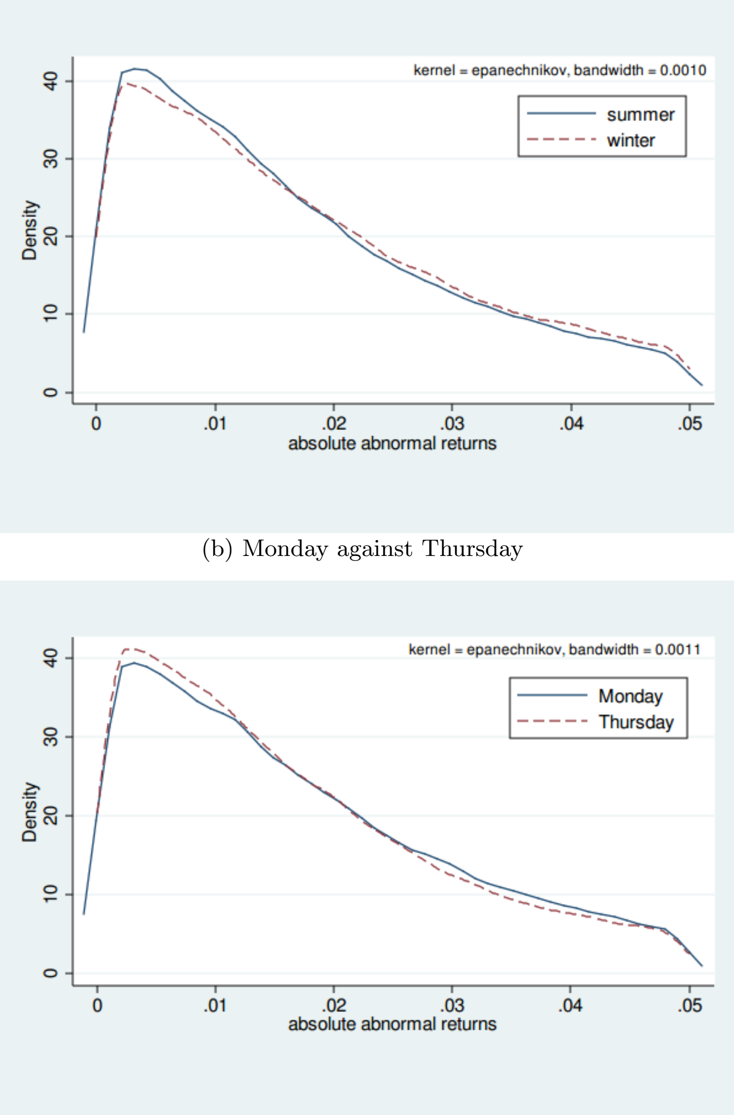
Trading activity shows patterns of substantial seasonality. The mean of absolute abnormal returns is smaller in the summer, .0219, while it is .0229 for the rest of the year. Similarly for weekdays, the mean is smaller for Thursday, .0219, than for Monday, .0235. Figure 2.3 shows conditional kernel density estimates.

The link between absolute abnormal returns and trading volume is not direct. Mean volume is higher on Thursday, 325,513 units against 291,138 units on Monday. But conditional volume averages show a little different picture for seasons of the year. Mean volume is lower in the summer, with 308,129 units against 321,368 units for the winter. So volatility as captured by absolute abnormal returns is larger in the winter, but so is the volume of trade.

In conclusion, absolute abnormal returns exhibit some serial dependence which is also confirmed by Wooldridge's test for first-order serial correlation (p-value $< .001$). In the analysis, this serial dependence ought to be accounted for.

Figure 2.3: Conditional densities of absolute abnormal returns

(a) Winter against summer



(b) Monday against Thursday

Note: The plots show absolute abnormal returns up to the 90th percentile. Trading activity shows different patterns in the winter and in the summer as well as at the beginning and at the end of the week. Stock returns are closer to market returns in the summer and on Thursday.

Chapter 3

Stock price effects of media mentions

With the data available, I'm about to test three theories about stock markets and the media. One is that of efficient markets. Under efficient markets, events get integrated into stock prices soon after they happen. How soon the soon is varies by the strength of the efficiency statement. Now I test the hypothesis that integration happens on the event day. Based on the data, I argue that this is not true.

The other theory is that of media impact on stock markets. The hypothesis is that the media does have a causal impact on stock trading. Evidence for this has been found by Tetlock (2011) and Shabani (2011) as well as Dougal et al. (2012) and Engelberg and Parsons (2011). I am not able to prove that the hypothesis is true but I show that the data does not falsify it, in fact it suggests it.

A variant of the hypothesis of media impact says that foreign news has an impact on trading activity in the country. Due to the small number of Financial Times mentions in the sample, statistical evidence is weaker than in the case of Indian mentions. Yet it only increases uncertainty about the magnitude, not the existence of the association between British mentions and the Indian stock market. It is clear from the data that price behaviour is different around Financial Times mentions. Furthermore, it is only different if they are not preceded by Indian mentions.

Whether this is merely due to Indian investors reading both newspapers, or due to foreign investors playing a role as well, is up to speculation.

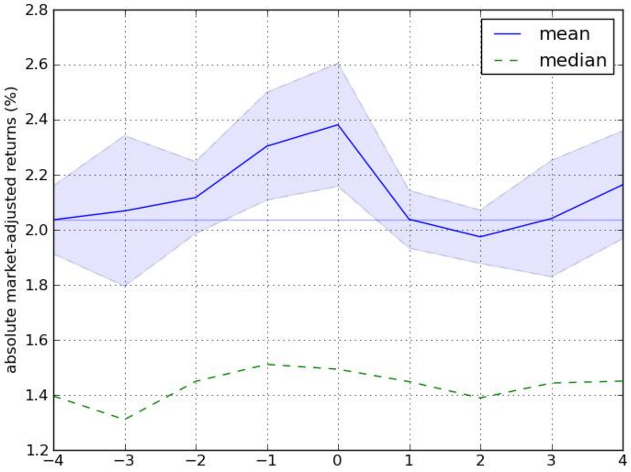
3.1 Volatility

A look at single mentions outlines the effect that Indian and British mentions seem to carry. Figure 3.1 shows daily averages of absolute abnormal returns for both the Economic Times and the Financial Times. In all of these cases, a single mention occurrs on day zero. Regressions presented in Table 3.1 confirm that the difference in trading activity is statistically larger only in the case of Indian mentions. The day right before the mention and on the day of the mention, divergence from market returns was about 0.3 percent points larger than usual. Divergence is back to normal after the mention.

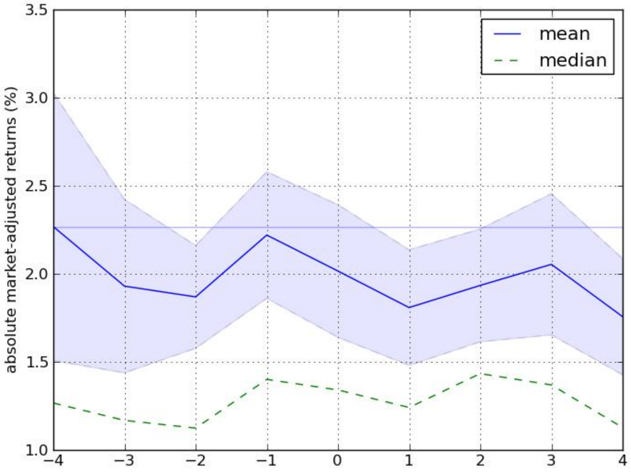
This suggests that investors to a significant degree obtain their price-relevant information from sources other than the financial media. The fact that there is no difference after the mention gives way to two interpretations. One is that investors without inside information still read the Economic Times and fully incorporate the impact of the event in the price. The other is that the timing is only due to chance, and investors would incorporate the event impact regardless of media mentions. To decide which interpretation matches reality better, events should be picked up based not on media mentions but on some exogenous factor,

Figure 3.1: Trading activity around a single mention

(a) Economic Times (India)



(b) Financial Times (UK)



Note: The figures show daily averages of absolute abnormal returns. A single mention occurs on day zero. For Indian mentions, activity is larger before and on the mention day. Table 3.1 shows that this remains with controls, too. For British mentions, volatility doesn't sway from usual levels.

Table 3.1: Trading activity around a single mention

	Economic Times (India)		Financial Times (UK)	
intercept	0.0204^{***}	0.0204^{***}	0.0227^{***}	0.0228^{***}
	(0.000630)	(0.000567)	(0.00388)	(0.00304)
day −3	0.000323	0.000313	−0.00338	−0.00337
	(0.00153)	(0.00146)	(0.00462)	(0.00233)
day −2	0.000807	0.000797	−0.00399	−0.00398
	(0.000918)	(0.000834)	(0.00416)	(0.00376)
day −1	0.00267^{*}	0.00265^{**}	−0.000481	−0.000466
	(0.00118)	(0.000820)	(0.00429)	(0.00422)
day 0	0.00345^{**}	0.00344^{**}	−0.00252	−0.00251
	(0.00130)	(0.00127)	(0.00433)	(0.00381)
day 1	0.0000191	−0.0000735	−0.00460	−0.00469
	(0.000826)	(0.000771)	(0.00423)	(0.00392)
day 2	−0.000619	−0.000686	−0.00334	−0.00367
	(0.000800)	(0.000742)	(0.00421)	(0.00368)
day 3	0.0000448	−0.0000840	−0.00214	−0.00260
	(0.00125)	(0.00122)	(0.00439)	(0.00392)
day 4	0.00127	0.00114	−0.00511	−0.00564
	(0.00118)	(0.00116)	(0.00423)	(0.00419)
mention fixed effects	NO	YES	NO	YES
observations	15,765	15,765	1,492	1,492

Note: See also Figure 3.1. The dependent variable is absolute abnormal returns. Mention happens on day zero. Coefficients show daily averages around the mention, relative to day −4. Trading activity is significantly higher before and on the mention day for Indian mentions. British mentions are statistically insignificant. Clustered standard errors in parentheses. Significance codes: $(^{*})$ $p < .05$, $(^{**})$ $p < .01$, $(^{***})$ $p < .001$.

say, corporate announcements. However, not having such data, I'm not going to test this now.

Financial Times mentions, on the other hand, are not in any way related to absolute abnormal returns. This is on one hand due to mentions that are likely missing. In the sample, the number of Financial Times mentions was about one tenth the number of Economic Times mentions. This naturally pushes the standard error of estimates upwards. However, it is also likely that the Financial Times mentions Indian companies much less often. This puts a natural bound on the accuracy of estimates on effects of British mentions that could be achieved.

Figure 3.1 and the corresponding table, Table 3.1, were generated in the following way. I collected every mention with no mentions in the preceding and the following four days. I groupped data up around these mentions, forming a panel with mentions along one dimension and trading days along the other. The equation to estimate was

$$|AR_{it}| = \alpha + \beta_t + \lambda_i + \varepsilon_{it},$$

with the constraint $\beta_{-4} = 0$ as this day was chosen as the base level in the regression. The coefficients were computed separately for Indian and for British mentions.

The data is not rich enough to analyse multiple-mention firms. There were only about eighty cases in the sample when mentions were followed by a single subsequent mention in two or three days. Statistical inference would be possible with data containing more mentions.

Table 3.2: Return momentum upon mentions

	day $t-1$ to day t	day t to day $t+1$
Economic Times (India)		
all	.0843*	−.0138
pure	.0922*	−.0312
single	.1020*	.0644*
Financial Times (UK)		
all	.0902	−.2365*
pure	.2206*	−.0408
single	.0777	−.1407

Note: The numbers are correlation coefficients. Mention happens on day t. “All” is the empire sample, “pure” contains only mentions that were not preceded by mentions in the other paper, “single” contains only pure mentions with no other mentions around. For comparison, without any mentions $r = -.037$ (p-value $< .001$). The symbol (*) marks significance at the 5 percent level.

3.2 Return momentum

Absolute abnormal returns mask the sign of the price change. However, the sign might also be informative. It turns out that around mentions, abnormal returns tend to have the same sign across days. This is in contrast with non-mention days when abnormal returns exhibit negative correlation. I conduct different tests to study this phenomenon.

3.2.1 A simple test of correlation

A simple non-parametric test of correlation shows that abnormal returns around mentions are indeed positively correlated. The test works with Pearson’s r, and is carried out as the following:

1. Pick every (AR_{it}, AR_{it+1}) pair such that firm i was mentioned on trading day $t+1$. Index these pairs by k as (AR_{k0}, AR_{k1}).

2. Compute

$$r \equiv \frac{\sum_{k=1}^{n}(AR_{k0} - \overline{AR_0})(AR_{k1} - \overline{AR_1})}{\sqrt{\sum_{k=1}^{n}(AR_{k0} - \overline{AR_0})^2}\sqrt{\sum_{k=1}^{n}(AR_{k1} - \overline{AR_1})^2}},$$

   where $\overline{AR_0}$ and $\overline{AR_1}$ are the means of these sequences.

3. From the original sample $\{AR_{k0}\}_k$, draw n abnormal returns with replacement, constructing $\{AR^*_{k0}\}_k$. Do the same for $\{AR_{k1}\}_k$, and compute r^*_b on the sample thus generated.

4. Repeat (3) $B - 1$ times.

B can be set to a thousand or ten thousand, according to convenience. The non-parametric p-value for r is obtained from the statistics $\{r^*_b\}_b$ computed with resampling. Since r is an ordinary correlation coefficient, it is straightforward to interpret. The sign is especially informative.

Correlation coefficients are summarised in Table 3.2. These coefficients were computed with the exclusion of outliers from the sample. To illustrate why this is crucial, take the lagged response to the Financial Times on the entire sample. It is $r = -.2365$ (p-value = .006) in the table. However, with the inclusion of a single outlier, Satyam Computer Services on January 7, 2009, the coefficient is $r = .3314$ (p-value = .002), that is, the sign switches and the result is similarly strongly significant. Looking at the news, it becomes evident how this one observation can drive the return momentum observed in the raw sample. The Financial Times published the following by the end of the trading day on January 7:

> “Indian shares fell by more than five per cent on Monday, despite a broader Asia Pacific rally to two-month highs. Satyam Computer Services shares plunged by around 80 per cent in Mumbai after its

> chairman confessed to fixing the IT outsourcing company's books for the past 'several' years. [...]"

This is not only an extreme event but the mention here is also endogenous. The market didn't react to the Financial Times, instead the Financial Times wrote about the company because of the market reaction. Accordingly, the abnormal returns of Satyam Computer Sevices on this day is -1.43, more than 28 standard deviations away from the average. From the present perspective, such events and reporting are obviously outliers.

As a benchmark, for pairs of days in the 2011–2012 sample when there were no mentions at all, $r = -.037$ (p-value $< .001$). This means that the Indian stock market features daily return reversal. This reversal is not large in magnitude but very statistically significant.

In the entire sample, for pairs of days with mentions on the second, $r = .035$ (p-value $= .02$). With only Indian mentions, $r = .084$ (p-value $< .001$), and with only British, $r = .090$ (p-value $= .063$). In the benchmark case, positive returns are followed by negative returns. With any kind of mention, however, returns tend to move on mention days in the direction they moved the day before. The relationship is statistically the weakest for British mentions, it is not significant at the five percent level.

It is important that the coefficient for British mentions increases and turns significant if mentions closely following Indian mentions are dropped. If the sample is thus restricted, I get $r = .2206$ (p-value $= .01$). Although the Financial Times data is severely lacking in quality, this result is evidence that price reaction around British mentions is different if there were also Indian mentions not much earlier.

This result is again reverted when only those British mentions are considered that are not preceded by and are not following other British or Indian mentions.

Then $r = .0777$ (p-value = .230) and it is insignificant once more. It is likely because multiple mentions that are dropped this way are associated with events of much higher impact. The coefficient can also be lower if there is a lagged price effect. In the case of multiple mentions, these coefficients pick up both the lagged effect of mentions on day $t - 1$ and the contemporaneous effect of mentions on day t. With single mentions, this is not so anymore.

I also assess lagged price effects. Correlation coefficients capturing these are shown in the last column of Table 3.2. In almost every case, they are statistically non-negative, and so they are different from the non-mention days benchmark. The only exception is British mentions when they can be preceded by Indian mentions. This suggests that information diffusion is not completed on the mention day but is more gradual.

3.2.2 Regression evidence

I run several regressions to test return reversal and return momentum. They are in line with the correlation coefficients, and confirm that there is positive dependence between returns around mentions. The complete specification of the regressions is

$$\begin{aligned} AR_{it} = \alpha + \rho AR_{it-1} + & \\ & + \beta_0 ET_{it} + \beta_1 ET_{it} AR_{it-1} + \\ & + \gamma_0 FT_{it} + \gamma_1 FT_{it} AR_{it-1} + \\ & + \boldsymbol{z_{it}}'\boldsymbol{\delta} + \lambda_i + \varepsilon_{it}, \end{aligned} \tag{3.1}$$

where λ_i is the firm fixed effect and $\boldsymbol{z_{it}}$ includes control variables.

Table 3.3 summarises the results. Lagged abnormal returns have a negative effect on contemporary abnormal returns. This is consistent with the correla-

Table 3.3: Price momentum upon mentions

	(1)	(2)	(3)	(4)
abnormal returns	-0.119***	-0.119***	-0.135***	-0.135***
	(0.0297)	(0.0254)	(0.0291)	(0.0291)
Economic Times	0.00125	0.00132	0.00104	
	(0.000856)	(0.000996)	(0.00102)	
ET×lagged AR	0.155***	0.155***	0.167***	
	(0.0384)	(0.0388)	(0.0402)	
Financial Times	-0.00469	-0.00476	-0.00537	
	(0.00544)	(0.00298)	(0.00302)	
FT×lagged AR	0.140*	0.140***	0.152***	
	(0.0590)	(0.0268)	(0.0292)	
pure ET				0.00129
				(0.00105)
pure ET×lagged AR				0.143***
				(0.0338)
pure FT				-0.00839
				(0.00432)
pure FT×lagged AR				0.307***
				(0.0899)
lagged avg. abs. AR			-0.0476***	-0.0483***
			(0.0131)	(0.0131)
lagged avg. AR			-0.0661**	-0.0653**
			(0.0204)	(0.0204)
lagged avg. illiquidity			0.0000269	0.0000271
			(0.0000346)	(0.0000347)
intercept	-0.000672***	-0.000673***	0.000336	0.000350
	(0.0000576)	(0.0000167)	(0.000292)	(0.000293)
firm fixed effects	NO	YES	YES	YES
observations	617,022	617,022	545,088	545,088

Note: Dependent variable is abnormal returns. "Pure" means mention is not preceded by mentions in the other newspaper. The basic relationship between abnormal returns today and yesterday is negative. It is statistically zero for Indian mentions. It turns positive for British mentions if not preceded by Indian ones. Clustered standard errors in parentheses. Significance codes: (*) $p < .05$, (**) $p < .01$, (***) $p < .001$.

tion obtained for non-mention trading days. On the other hand, if there are mentions in the Economic Times or in the Financial Times, the overall effect of lagged abnormal returns turns non-negative—again, consistent with the positive correlation coefficients found around mention days. Moreover, the overall effect is statistically positive for Financial Times mentions if they are not preceded by Economic Times mentions.

3.3 Textual analysis

A promising direction in better capturing how news are related to each other is introducing textual similarity. Tetlock (2011) has textual analysis as the cornerstone of his study, defining staleness as an average of pairwise similarity indices. I present only a very simple attempt at exploring news similarities in my sample. The topic is elaborated on in more detail and with more sophistication by Tetlock et al. (2008).

A measure of similarity is important to capture staleness of the news. British mentions could be sheer reiterations of information already published in India, while counting as new in Britain. If the market reacts to similarity with previous Indian news, it gives ground to an interpretation of investor attention.

A popular way of assessing textual similarity is generating what are called n-grams, and comparing those. Unigrams are individual words that occur in the text. Bigrams are pairs of adjacent words. To illustrate, the sentence *"Send toast to ten tense stout saints' ten tall tents"* would have the corresponding unigrams *send, toast,* etc., and bigrams *(send, toast), (toast, to),* and so on.[1]

Further, n-grams are usually not generated on unprocessed text. First, cer-

[1] It is customary to include the first word on its own as well. In this case this would be *(·, send).*

tain words called *stopwords* that only have grammatical roles but carry little meaning are dropped. Second, the remaining words are stemmed, e.g., *hiked* becomes *hike* and *pictured* becomes *pictur*.

I use a different method, carrying out a grammatical analysis of every sentence, and keeping only nouns and verbs in their lemmatised forms. Lemmatisation is similar to stemming but produces a form of the word that makes sense in itself. To provide an example, with lemmatisation *hiked* translates to *hike* the same way, but *pictured* becomes *picture* instead. Lemmatisation is computationally more intensive, but is more straightforward to implement as it doesn't require a list of somewhat arbitrarily defined stopwords. Of the resulting lemmas, I only keep verbs and nouns, and ignore the rest.

Using the lemmas thus extracted, I pick news mentioning the same company and, following Tetlock (2011), compute a similarity measure between news i and news j as

$$S_{ij} = \frac{\#\left(L_i \cap L_j\right)}{\#\left(L_i \cup L_j\right)}.$$

Here, $\#(\cdot)$ denotes set cardinality, and L_i is the set of lemmas collected from news i. This formulation is also called the Jaccard index. The final similarity measure of a mention is the average of its similarity with all mentions on the preceding five trading days.

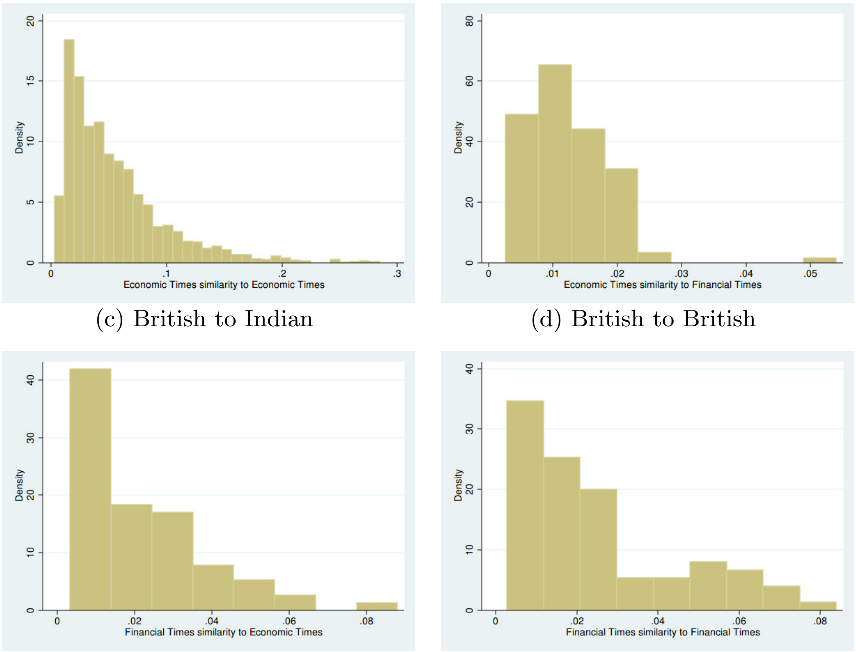
The indices thus generated are plotted on histograms in Figure 3.2. Unfortunately the sample size is too small for all but the Indian to Indian direction to provide useful reference regarding staleness. For Indian mention similarity to Indian mentions, the measured effect is minuscule.

One sign that the attempt at making use of textual analysis is heading in the right direction is shown by a Table 3.3-like regression. I replaced ET$\times$lagged AR with ET$\times$(1$-$ET similarity to ET)$\times$lagged AR. This more complicated variable

Figure 3.2: Histograms of the textual similarity index

(a) Indian to Indian

(b) Indian to British



(c) British to Indian

(d) British to British

Note: The histograms show the distribution of non-zero values for the similarity index. The shape of the distributions appears to be similar. Sample sizes are (a) 2,187, (b) 119, (c) 72 and (d) 83. The sample being small for all cases but (a), measured similarity tends to be low.

takes the value zero if there was no mention, one if there was a mention and it was not similar to a previous mention, and values inbetween if there were other, similar mentions. The coefficient estimate basically remains unchanged, but the standard errors are much smaller. This indicates better fit and shows that the construction of the similarity index is not mistaken. However, it might need further testing on more data and improvements to perform really well.

3.4 The lessons learnt

I evaluate three hypotheses on my data. The first says that markets are efficient. Figure 3.1 and Table 3.1 show that stock returns are significantly more volatile for two days around the mention (on the mention day and one day before that). Furthermore, Table 3.2 documents return momentum to the mention day from the day before. This finding is supported by regressions shown in Table 3.3.

These findings could be explained under efficient markets by saying that in such cases there were in fact two independent underlying events. The market moves in a predictable direction on the mention day because there was an event the day before and on the mention day as well. Having said this, it would also require that the events be both positive or both negative, systematically. This explanation is quite unlikely.

It also appears that news are not absorbed immediately by the market. Information diffusion takes longer than one trading day. Although return momentum is not observed from the mention day to the day after, neither is return reversal which is the market pattern in the absence of mentions.

The second hypothesis says that the media has an impact on markets. Lacking exogenous variation in media coverage, I cannot prove this statement. Nevertheless, the data doesn't disprove it, either. Return momentum is measured to be the largest on the day of the mention, and vanishes the day after. This is consistent with the story of media impact on markets.

The third hypothesis is a variation of this. It says that foreign media has an impact on domestic markets. This is harder to measure, there being so little data on Financial Times mentions, but even when there is no statistically significant price momentum, return reversal is still not observed. Reversal is seen only with the British mention coming after Indian mentions. This indicates that the Financial Times is read by investors who are present in Indian stock trading.

Whether this is simply because Indian investors also read the Financial Times, or rather because there is a sizeable group of foreign investors who read the Economic Times, too, is not clear and is open to speculation. As one point in this inquiry, note that information diffusion, i.e., return momentum, spanning beyond the day of the mention means that a significant proportion of investors doesn't follow that newspaper. (All this is of course assuming that there is no new event the day after.) And as Table 3.2 shows, return reversal doesn't kick in the day after the mention for either the Economic Times (significant return momentum) or the Financial Times (no relationship between returns). This suggests that both newspapers have a significant "non-reader" base.

Regarding foreign investors in India, Griffin et al. (2007) document several facts about the Indian stock market. India was turning over 1.31 percent of the market in one week which is below the 1.48 percent of the United Kingdom and much below the 2.26 percent of the United States. This measure can be interpreted as a weight indicator of small investors. Wealth distribution in India might confine the stock market to be relevant for only a relatively thin layer of society, and this thin layer might get their information from sources other than newspapers. Another interesting statistical figure is trading volume by foreigners which was only 10 percent of the average trading volume. This is lower in India than for instance in Indonesia, Japan, or South Korea. This could be one explanation why the Financial Times doesn't have as clear an association with returns as the Economic Times.

In conclusion, what is indispensable in getting a better picture of foreign media impact is more data on Financial Times mentions.

Chapter 4

Conclusion

In this thesis, I tested three hypotheses on stock markets and media coverage. The first is that of efficient markets which I found difficult to defend against the data. The second is that of media impact on stock markets. Not having exogenous variation in media coverage, I cannot provide conclusive evidence. Yet the hypothesis is suggested by several of the statistical tests I conduct. The third hypothesis is a variant on this, saying that foreign media has an impact on domestic markets. In so far as possible given constraints of the data, media impact is indicated in this case as well. Daily return momentum is significant on mention days, and immeasurable the day after.

To improve on the inference made possible here, better data is necessary on news, specifically company mentions. The web crawling procedure that I used for this thesis does not appear to be efficient enough for discovering a sufficient number of mentions in the Financial Times. More resources spent on crawling, or direct database access to Financial Times news would both be great help in solving this problem. Better data on British mentions could give way to more sophisticated approaches, including the analysis of textual similarity of news reporting, to proving or disproving the hypotheses discussed.

Bibliography

Dougal, Casey, Joseph Engelberg, Diego García, and Christopher A. Parsons, "Journalists and the Stock Market," *The Review of Financial Studies*, 2012, *25* (3), 639–679.

Engelberg, Joseph E. and Cristopher A. Parsons, "The Causal Impact of Media in Financial Markets," *The Journal of Finance*, 2011, *66* (1), 67–97.

Griffin, John M., Federico Nardari, and Ren M. Stulz, "Do Investors Trade More When Stocks Have Performed Well? Evidence from 46 Countries," *The Review of Financial Studies*, 2007, *20* (3), 905–951.

Shabani, Reza, "Corporate News, Asset Prices, and the Media," 2011. Job market paper.

Tetlock, Paul C., "All the News That's Fit to Reprint: Do Investors React to Stale Information?," *The Review of Financial Studies*, 2011, *24* (5), 1481–1512.

_ , Maytal Saar-Tsechansky, and Sofus Macskassy, "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *The Journal of Finance*, 2008, *63* (3), 1437–1467.