News, asset prices and capital flows: Evidence from a small open economy

Galen Sher*

January 20, 2017

Abstract

I present evidence from South Africa that domestic asset prices and capital flows between residents and non-residents reflect the content of domestic print news media. I find that the contents of national newspapers can predict 9% of the variation in daily stock returns one day ahead and 7% of the variation in the daily term premium three days ahead. This predictability in stocks and bonds coincides with predictability of the content of domestic print news media for net equity and debt portfolio capital inflows, suggesting that the domestic print news media affects foreign residents' demand for domestic assets. Moreover, predictability of domestic print news media for near future stock returns is driven by emotive language, suggesting a role for 'sentiment', while such predictability for stock returns further ahead and the term premium is driven by non-emotive language, suggesting a role for other media factors in determining asset prices. These results do not seem to reflect a purely historical phenomenon, finite-sample biases, reverse causality, serial correlation, volatility or day-of-the-week effects. The results support models where foreign agents' short-run beliefs or preferences respond to the content of domestic print news media heterogeneously from those of domestic agents, while becoming more homogeneous in the medium term.

Keywords: Asset pricing, capital flows, machine learning
JEL Classification: G12, G15, G17, C58

1 Introduction

In the classic representative agent closed economy consumption asset pricing model, risky asset prices are determined by preferences over rates of time preference and risk aversion, and beliefs about expected future returns and volatility. Investor psychology therefore plays an integral role in asset pricing. In this context, news in the form of changes in preferences or beliefs should affect asset prices.

Campbell et al. [1993] formalise this intuition in the context of a heterogeneous agent model, where some agents have uncertain future levels of risk aversion and other agents cannot fully offset these agents' buying or selling pressure because they themselves are risk averse. News in the form of changes in the first group's risk aversion can therefore affect risky asset prices, even when agents are rational in the sense that their beliefs are consistent with the model of the economy in which they operate. However, when some agents have beliefs about future returns or volatility that are irrational, as in the model of De Long et al. [1990], and when other agents have constraints like finite investment horizons or risk aversion that prevent them from offsetting the buying or selling pressure of irrational agents, news in the form of changes in the beliefs of irrational agents can also affect risky asset prices.

*I would like to thank my advisor Dave Strugnell for helpful comments on many drafts of this paper. Email gsher@imf.org.

Several papers study the role for news in determining closed economy stock prices by evaluating the extent to which observed measures of news can explain, in a statistical sense, variation in observed stock returns. In an early work, Cutler et al. [1988] find that at most one third of the daily variation in stock prices can be accounted for by the unexpected component of contemporaneous variation in several macroeconomic indicators. The authors conclude that stock prices must be determined by factors other than future cashflows or discount rates, which suggests an important role for preferences and beliefs.

Two closely related papers, Tetlock [2007] and Garcia [2013], study the role for 'sentiment' in determining stock prices. Using data for the period 1984–1999, Tetlock [2007] finds that the words appearing in the Wall Street Journal's "Abreast of the Market" column on a given morning provide a leading indicator of the change in the Dow Jones Industrial Average Index on that day. Specifically, Tetlock [2007] defines a single media factor as the difference between the numbers of emotionally 'negative' and 'positive' words, according to the Harvard IV-4 dictionary, which appear on each day and bases his findings on in-sample hypothesis tests. Since negative word counts are negatively associated with that day's stock return, whether measured between consecutive days' closing prices or between 10 a.m. and the time of market close on the same day, and since the association between negative word counts and future day's stock returns reverses sign at longer forecast horizons, the author concludes that there is some evidence that stock prices reflect media sentiment rather than information about future cashflows or discount rates. One uncertainty in this conclusion is the potential role for finite sample bias, given the large number of parameters being estimated relative to the number of observations.[1] Garcia [2013] applies the method of Tetlock [2007] to find a similar association between word counts from articles published in the New York Times on a given day and changes in the Dow Jones Industrial Average on that day, over the 1905–2005 period. The author also documents a procyclicality in the magnitude of this relationship. The large number of observations in Garcia [2013] limits concerns about finite sample biases present in Tetlock [2007]. However, two explanations acting together could account for the results in Garcia [2013]. The predictability of news for changes in stock prices could reflect a pre-WWII phenomenon and it could reflect reverse causality from financial market developments after the NASDAQ closed at 4 p.m. to the content of news articles published the next morning. The author investigates each potential cause in isolation, and the results weaken appreciably after allowing for the second potential explanation, but the author does not allow for both explanations together.[2,3]

Other authors have investigated the informativeness of other news sources for stock returns. Antweiler and Frank [2004] find that the bullishness of messages on stock message boards provide only limited information about future stock returns of individual companies. Bullishness on a given day is an aggregate of the bullishness of each individual message that was posted on that day, and the bullishness of each individual message is measured using a statistical algorithm to extrapolate from subjective bullishness ratings for a sample of messages. The extrapolation procedure is based on frequencies of occurrence of individual words. The authors also find that disagreement between the messages is informative about trading volume, and that the number and bullishness of messages are both informative about future volatility. Luss and d'Aspremont [2015] find that individual company press releases are informative about future intraday company-specific stock volatility, but not about the direction of future company-specific stock returns. The authors' word list is ad hoc, but has the

[1] Specifically, each of the three specifications being estimated contain 27 coefficients and 5 Newey–West standard errors to be estimated. The specification for trading volume contains an additional 5 coefficients. This gives 101 parameters to be estimated using $3,709$ observations, or 37 observations per parameter.

[2] Garcia [2013] investigates the reverse causality explanation by measuring the association between a given day's word counts and the stock return between 11 a.m. and the time of market close on that day. However, the author's sample for such changes in intraday stock prices goes back to 1933, so that the measured association could still be reflecting a historical phenomenon.

[3] There is also uncertainty from a point of inconsistency between Garcia [2013] and Loughran and McDonald [2011]. The latter authors find problems with applying the Harvard IV-4 dictionary used in Tetlock [2007] to documents covering financial news, where the vocabulary differs from standard English. This leads them to propose an alternative dictionary of emotionally 'positive' and 'negative' words that is more suited to financial applications and claim that it produces materially different results to those of the Harvard IV-4 dictionary. Garcia [2013] uses the dictionary of Loughran and McDonald [2011], but claims to produce results that are qualitatively similar to those of Tetlock [2007].

advantage of containing some pairs of words like *didn't increase*. Da et al. [2015] find that the volume of internet search queries related to words like 'recession', 'bankruptcy' and 'unemployment' is informative about future stock return reversals, stock return volatility and equity mutual fund outflows. The list of individual words that the authors use to compute this volume on a given day is based on the Harvard IV-4 dictionary of emotively negative words, tailored to be related to the terms used in search queries and sampled from based on their covariance with stock returns.

These empirical studies specify word lists based on emotive words, which are suitable for studies of the role of sentiment in asset pricing, but leave open the role of other print news media factors in affecting asset prices. In particular, the emotive words commonly used ignore macroeconomic terms, which may be important for affecting agents' beliefs and preferences. The print news media can be informative for stock returns and can affect asset pricing without necessarily affecting sentiment. If other print media factors turned out to be important for asset pricing, the measured effect of the single print media factor would be affected by such omitted variables.

In an open economy asset pricing model, foreigners' preferences and beliefs determine their demand for domestic assets, which affects domestic asset prices. If foreigners are at an informational disadvantage relative to domestic residents about domestic assets, foreigners may rely more on such publicly available information as print news media to inform their beliefs about future returns and volatility of domestic assets, while domestic residents have access to additional sources. Hence, foreigners could behave like the irrational agents in the model of De Long et al. [1990] in having their beliefs determined by such publicly available information as print news media that does not necessarily provide information not already incorporated into asset prices. In the context of a small open economy like South Africa, where domestic asset values are small relative to those of the rest of the world and controls are limited, foreigners have the potential to influence domestic asset prices substantially.

Empirical literature on the role for news in determining capital flows is limited, even though the literature on capital flow surges and sudden stops makes qualitative references to the important role for "market sentiment". Fratzscher [2012] finds a role for news, in the sense of deviations in macroeconomic variables from median expectations expressed in Bloomberg surveys, in explaining capital flows. The macroeconomic variables considered, including the trade balance, gross domestic product, industrial production and unemployment, do not provide much information about changes in preferences or beliefs that are relevant for determining asset prices.

In this paper, I revisit the literature on the role for print news media in affecting asset prices and fill the gap in the literature on the role for news in affecting capital flows, using data from South Africa, which is a small open economy. Reproducing the analyses of Tetlock [2007] and Garcia [2013] on an archive of some fifteen thousand news articles published in South African national newspapers between 2008 and 2014, I cannot find evidence to support the role for their single print news media factor in explaining stock returns in sample. However, when I generalise the single factor representation of newspaper articles to a multi-factor representation, I find strong out-of-sample evidence supporting a role for the print news media in explaining stock prices. Using such a multi-factor representation of the content of newspaper articles, I find that 9 percent of the daily out-of-sample variation in stock prices can be accounted for by the content of newspaper articles, and these results do not reflect purely historical phenomena, finite-sample biases, omitted variable biases, reverse causality or proxying for other sources of predictability like lagged returns, volatility or day-of-the-week effects. This finding therefore eliminates the uncertainties identified above in the existing literature.

I explore the explanatory power of different types of language through different multi-factor representations of the content of print news media, which allows me to contrast emotive language with non-emotive language, and simple vocabularies based on individual words with complex vocabularies based on collections of words. This process allows me to assess whether the print news media affects asset prices through sentiment or other fundamental channels. I also explore the role for multi-factor representations of print news media in explaining bond prices and capital flows up to five days ahead.

The multi-factor representation of domestic print news media can explain out-of-sample variation in aggregate stock returns, aggregate stock trading volumes and net portfolio equity capital inflows one and two days ahead. This finding suggests that domestic print news media affects aggregate stock

prices through foreign demand with a lag of one or two days. This predictability also appears to be driven by emotive language, which suggests that the domestic print news media influences foreign demand through sentiment. The predictability holds out of sample, limiting concerns about finite sample biases, and after controlling for lagged returns, volatility and day-of-the-week effects, which limits concerns about proxying for these other potential sources of predictability.

Three and four days ahead, the multi-factor representation of domestic print news media similarly shows excess predictive power for aggregate stock returns, but not for trading volume or net equity portfolio capital inflows. This finding suggests that foreign and domestic agents' beliefs and preferences are affected by the content of domestic print news media three and four days prior, but that these two groups of agents respond similarly to such content. The predictability three and four days ahead is driven by more complex non-emotive vocabularies, suggesting that the response of beliefs and preferences to domestic print news media does not operate through sentiment.

In addition to its role in determining stock prices, I find that the multi-factor representation of print news media can explain 7 percent of the daily variation in the term premium three days ahead. This finding suggests that the domestic print news media affects the price of long-term domestic bonds relative to short-term domestic bonds with a lag of three days. The three day lag coincides with predictability of the domestic print news media for net portfolio debt capital inflows three days ahead, suggesting that foreign demand for domestic long-term bonds relative to short-term bonds depends on the domestic print news media three days prior. This predictability also appears to be driven by non-emotive language, which also suggests that the domestic print news media influences foreign demand for domestic long-term bonds without affecting foreign sentiment.

This paper is structured as follows. Sections 2, 3 and 4 describe the data sources, analytical methodology and results respectively. Section 5 summarises conclusions from the existing work and proposes further work for the coming months.

2 Data

The data for this project are text from published news articles, historical market index levels from the Johannesburg Stock Exchange (JSE) and historical capital flows. I construct an archive of 15,584 South African news articles published in The Times, Business Day and Financial Mail between 11 December 2008 and 6 February 2014.[4] I obtain these articles from Factiva, a news provider owned by Dow Jones & Company, with the search term "South African economy". The articles are time stamped with their day of publication, so text can be analysed at a daily or lower frequency.

These three newspapers are distributed in print nationally and are available free of charge online. All three are owned by Times Media Group, headquartered in Johannesburg. Between August and October 2016, the websites of The Times, Business Day and Financial Mail received 19.7, 1.2 and 1.1 million visits respectively.[5] Print editions of The Times and Business Day are published every weekday and those of the Financial Mail are published weekly on Fridays. According to figures released by the Audit Bureau of Circulations, between July and September 2016 the circulations of these three print publications were 59, 23 and 13 thousand copies per publication date respectively, down from 109, 26 and 15 thousand copies per publication date a year earlier.

I obtain minute-by-minute Top 40 futures contract prices from Portara CQG.[6] Open market trading of equity derivatives on the JSE takes place between 8:30 a.m. and 5:30 p.m. local time. The 8:30 a.m. opening price is determined by an auction conducted between 8:25 a.m. and 8:30 a.m.. This 8:30 a.m. price is therefore a post-open price in the sense that it reflects information released since the close of trading on the previous trading day. I calculate post-open stock returns using changes in the

[4] To the best of my knowledge, Factiva does not provide bulk downloads from their news database. It would therefore be difficult to extend the newspaper archive that I have.

[5] Data on numbers of website visits are provided by SimilarWeb Ltd, a digital market intelligence company.

[6] The minute-by-minute futures contract prices refer to Top 40 Index futures contracts traded on the JSE. These prices are back-adjusted as at 31 October 2016 and the contracts are rolled over one day before expiry to create a continuous time series of futures prices.

natural logarithm of the Top 40 index futures contract price between various combinations of starting and ending times.[7]

I obtain daily historical equity total return indices and trading volume of the JSE All Share Index and Top 40 Index from Bloomberg. From the same source, I obtain Bloomberg South Africa bond price indices for 10 and 1-3 year maturities.[8] I obtain the GOVI daily historical total return index of South African government bond prices from Thomson Reuters Datastream. This index is calculated and published by the JSE and consists of the ten largest and most liquid South African government bonds by market capitalisation and clean consideration turnover respectively. I calculate daily stock and bond returns as the change from one day to the next in the natural logarithm of the respective total return index. I compute daily trading volume as the change from one day to the next in the natural logarithm of the sum of one and trading volume. I calculate the daily equity premium as the difference between the daily stock and bond returns, and I calculate the term premium as the difference between the daily change in the natural logarithm of the 10 year maturity bond price index minus the daily change in the natural logarithm of the 1-3 year maturity bond price index.

I obtain daily net portfolio equity and debt capital flows into South Africa in US dollars from the Institute for International Finance. Although capital flow data are available for weekends, I only consider capital flows on trading days to ensure consistent samples for asset prices and capital flows. The daily asset prices and capital flows data cover the same period as the newspaper archive, while the intraday equity futures prices cover the period between 14 October 2010 and the last date of the newspaper archive.

3 Method

3.1 Matching news to returns

As explained in Section 2, I consider two sets of dependent variables. The first are intraday stock (futures) returns and the second are daily stock returns, bond returns, percentage changes in trading volume, equity premium, term premium and capital flows. I match each set of dependent variables slightly differently to articles published in the national newspapers. By tailoring the matching procedure to each set of dependent variables, I am able to ensure that I minimise any overlap between publication times of newspaper articles and market trading times.

Consider first the matching of newspaper articles to intraday stock returns. I match all articles published on the previous trading day and any intervening calendar days to the post-open stock return on any given day. In particular, I match newspaper articles published on Friday, Saturday and Sunday to the intraday stock return on Monday. I match newspaper articles published on Monday to the intraday stock return on Tuesday, and so on. This matching scheme avoids any potential for the content of the news articles to reflect the changes in stock prices that they are to be used to predict. I am able to match 815 days of intraday returns to newspaper articles in this way.

Next consider the matching of newspaper articles to the dependent variables observed at the daily frequency. For ease of exposition I explain with the example of daily close-to-close stock returns. If we enumerate the trading days on which we observe closing stock prices as $\ldots, t-1, t, t+1, t+2, \ldots$, then I match the stock return y_{t+1} between trading days numbered t and $t+1$ to those newspaper articles published on the date of the trading day numbered t and, if applicable, any non-trading calendar days earlier than the date of the day numbered t but (strictly) later than the date of the trading day numbered $t-1$. This means that I match newspaper articles published on Friday to

[7] Tetlock [2007] uses changes in the Dow Jones Industrial Average between 10 a.m. and 4 p.m., Garcia [2013] uses changes in this index between 11 a.m. and 4 p.m.. Trading on the NYSE and NASDAQ, which contain the stocks that make up the Dow Jones Industrial Average, takes place between 9:30 a.m. and 4 p.m.. Trading on the JSE takes place between 8:30 a.m. and 5:30 p.m.. The JSE has an auction period between 8:25 and 8:30 a.m. and an administration period for allocations and reporting between 5:30 and 6:15 p.m..

[8] The Bloomberg tickers for these securities are `JALSH Index`, `TOP40 Index`, `BSAFR10 Index` and `BSAFR13 Index` respectively.

the close-to-close stock return computed between Friday close and Monday close. It also means that I match newspaper articles published on Saturday, Sunday and Monday to the close-to-close stock return computed between Monday close and Tuesday close. I am able to match 1283 daily stock returns to newspaper publication days in this way. In what follows, I also consider the ability of print news media to explain stock returns up to 5 days ahead. The matching procedure is analogous to the one-step ahead case, replacing y_{t+1} by y_{t+n} for $n = 1, 2, 3, 4, 5$.

Note that I match newspaper articles published on Sunday to Monday's post-open stock return and the daily return between Monday close and Tuesday close. This matching procedure ensures that in each case, newspaper articles published on Sunday are matched to the nearest available future return that does not coincide with Sunday. A procedure of matching Sunday's newspaper articles to daily stock returns between Friday close and Monday close, for example, would allow for the possibility that any financial market developments between the close of the domestic stock market on Friday and the publication of a newspaper article on Sunday could influence the content of that newspaper article. I would like to rule out such reverse causality interpretations of any role for print news media in the pricing of assets.[9] Although this matching procedure minimises the potential for reverse causality in the case of one-step ahead daily returns, it does not eliminate this interpretation in this case because a newspaper article time stamped with a given day may have been published after the close of trading on that day. This potential for reverse causality is the motivation for considering post-open stock returns, where such an interpretation is precluded. Nevertheless, the results for intraday and one-step ahead daily returns are similar, which shows that the reverse causality explanation is not important in the case of one-step ahead daily returns.[10]

I then convert the text of these matched articles into numeric features. If multiple articles are matched to a given stock return, the bodies of those articles are concatenated into one document, so that I end up with a time series of documents for the purposes of extracting features. I convert each document to numerical features based on word lists that I refer to as *vocabularies*. A vocabulary could be a list of individual words, a list of individual words and pairs of consecutive words, or a list of individual words, pairs of consecutive words and triples of consecutive words. Three vocabularies that suggest themselves are the positive, negative and combined positive and negative individual words as defined in Loughran and McDonald [2011] and used in Garcia [2013]. Another natural vocabulary is the union of all individual words occurring in any of the articles.[11] To allow for terms with qualifying words like *not good*, I also consider an extension of the preceding vocabulary that includes all individual words and all pairs of consecutive words occurring in any of the articles. Finally, to allow for terms with qualifying words that may be separated from the word they qualify, like *not very good*, I consider an additional vocabulary based on all individual words, all pairs of consecutive words and all triples of consecutive words appearing in any of the articles. I therefore end up with six vocabularies, each one of which will produce a set of explanatory variables that can be used in a predictive regression model.

Following Luss and d'Aspremont [2015], I calculate term frequency–inverse document frequency (TF-IDF) features from these vocabularies. The TF-IDF feature for each vocabulary item (i.e. an individual word, word pair or word triple) in a given document is TF $\times \log(m/\text{DF})$ where TF is the number of occurrences of this item in this document, m is the total number of documents and DF is the number of documents in which this item appears.[12] For a given document, this procedure produces one feature number for each item in a vocabulary so that each document is represented by a vector. For vocabularies with many items, these vectors can be long. The Loughran and McDonald [2011] dictionaries contain 354 positive words and 2355 negative words, and the collection of all articles

[9] The JSE closes for stock trading at 5 p.m. according to the Coordinated Universal Time + 2 hours time zone, which is 6 hours before the New York Stock Exchange closes, at 4 p.m. according to the Coordinated Universal Time - 5 hours time zone.

[10] The reverse causality interpretation also does not apply to the relationship between print news media and n-step ahead daily returns for $n > 1$.

[11] This vocabulary is called the bag-of-words model in computational linguistics.

[12] The TF-IDF features therefore downweight vocabulary items that occur many times in across all newspaper articles, because these items are less likely to be useful for discriminating between documents associated with high and low stock returns.

contains 67 thousand individual words, 1.6 million individual words and pairs of consecutive words, and 5.8 million individual words, pairs of consecutive words and triples of consecutive words. Each observed stock return is therefore associated with exactly one document, which is a collection of all the newspaper articles matched to that stock return, and each document is transformed into a vector of up to 5.8 million numeric features that describe its content. These stock returns and numeric features can then be used to train a prediction model.[13] I explain the model specifications in Section 3.2 and estimation procedure in Section 3.3.

3.2 Model specification

I consider the role for print media factors in explaining stock returns, bond returns, the equity premium, the term premium, stock trading volume and capital flows. For each dependent variable, I consider models with print media explanatory variables only, with a set of non-print media explanatory variables only and with a combined set of print media and non-print media explanatory variables jointly. For the first four dependent variables, I use five lags of the dependent variable, five lags of the squared dependent variable and day-of-the-week indicator variables as non-print media explanatory variables. For the stock return dependent variable, these explanatory variables match those used in Tetlock [2007] and Garcia [2013].[14] When stock trading volume, equity capital flow or total equity and debt capital flow are the dependent variable, I use five lags of stock returns, five lags of squared stock returns and day-of-the-week indicator variables as non-print media explanatory variables. When debt capital flow is the dependent variable, I use five lags of bond returns, five lags of squared bond returns and day-of-the-week indicator variables as non-print media explanatory variables.

3.3 Model estimation

For predicting stock returns based on text features, I use a machine learning technique called support vector regression (SVR) due to Vapnik [1995]. This technique has been found to have good performance in the context of predicting future volatility of individual stock returns based on individual word and pair of consecutive words TF-IDF features extracted from companies' own 10-K filings to the US Securities and Exchange Commission [Kogan et al., 2009]. Luss and d'Aspremont [2015] use an earlier version of this technique, which is designed for predicting binary response variables, to predict the sign of stock returns from news features.

A support vector regression predicts a scalar outcome y from a vector input x using the linear regression model $w \cdot \phi(x) + b$ where w is a vector of parameters, b is a scalar parameter and ϕ is a function. In this paper, I consider the simple case of linear SVR where $\phi(x) := x$ for all x. The parameters w, b must be estimated from a sample of data

$$(y_1, x_1), (y_2, x_2), \ldots, (y_l, x_l). \tag{1}$$

Among the many possibilities for estimating the parameters w, b on the basis of the sample (1), a SVR

[13] Note that we have more explanatory variables than observations, which makes the estimation of classical regression models like ordinary least squares infeasible, but other regression models that employ regularisation techniques remain feasible.

[14] Following Garcia [2013], the stock returns used to construct lagged explanatory variables follow the same definition as the stock returns used for the dependent variable. Therefore, for models where the dependent variable is the change in the natural logarithm of the futures price between 10:30 a.m. and 5:30 p.m., the lagged return explanatory variables would be calculated from these returns.

introduces the two free ('tuning') parameters ϵ, C and estimates w, b as the solutions to

$$\begin{gathered}\min_{w,b,\xi_1,\ldots,\xi_l,\xi_1^*,\ldots\xi_l^*} \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \\ \text{subject to } y_i - w \cdot x_i - b \leq \epsilon + \xi_i, \\ w \cdot x_i + b - y_i \leq \epsilon + \xi_i^* \\ \text{and } \xi_i, \xi_i^* \geq 0 \text{ for all } i = 1, \ldots, l. \end{gathered} \tag{2}$$

In this paper, I choose $\epsilon := 0.1$ and consider $C = 2^{-8}, 2^{-1}, 1$ and 2.

3.4 Performance assessment

I assess the performance of this model by computing out-of-sample prediction errors. In particular given a full sample (1) with size $l = m$, I consider fitting the model (2) on 100 expanding window subsamples of size $l < m$ for $l \in \{\lfloor \frac{m-1}{100} \rfloor l' : l' = 1, 2, \ldots, 100\}$ and obtaining each model forecast $\hat{y}_{l+1}$ given the input x_{l+1}. I then compute the out-of-sample cumulative root mean square prediction errors

$$\sqrt{\frac{1}{l'} \sum_{l''=1}^{l'} \left(y_{\lfloor \frac{m-1}{100} \rfloor l''+1} - \hat{y}_{\lfloor \frac{m-1}{100} \rfloor l''+1} \right)^2} \tag{3}$$

for each $l' = 1, 2, \ldots, 100$.

Against this model I also consider a naive benchmark returns forecasting model that predicts the next day's return will be the historical mean return observed up to the date at which the prediction is made. Replacing the term in (3) involving $\hat{y}$ by $\sum_{l=1}^{\tilde{l}} y_l / \tilde{l}$ where $\tilde{l} = \lfloor \frac{m-1}{100} \rfloor l''$, I obtain the cumulative root mean square prediction error for a benchmark model that forecasts the historical mean return. This benchmark model is agnostic about the future direction of returns above or below the historical mean, so outperforming this model indicates in particular that we tend to predict the deviations of future returns from the historical mean better than could be expected by pure chance. This historical mean benchmark model also has the virtue that its mean square is simply the variance of the dependent variable, so if a candidate model outperforms this benchmark model in terms of mean square prediction error, then the candidate model produces smaller errors on average than the variance of the returns being modelled. In the results presented below, I express the cumulative root mean square prediction error (3) as a fraction of the cumulative root mean square prediction error of the historical mean model. A ratio less than unity indicates that the candidate forecasting model outperforms the benchmark historical mean model in terms of out-of-sample mean square prediction error.[15]

4 Results

4.1 A single media factor

I estimate the parameters in the the specification

$$y_{t+1} = \alpha + \beta f_t + \sum_{k=1}^{5} \phi_k y_{t-k} + \sum_{k=1}^{5} \gamma_k y_{t-k}^2 + \theta_t + e_t \quad \text{for all } t \tag{4}$$

where y_t denotes the post-open stock return on trading day t, f_t is the single media factor and θ_t is a set of day-of-the-week indicator variables, by ordinary least squares. Following Tetlock [2007]

[15] I use the errors from the benchmark forecasting model as a device for rescaling the errors from the main candidate forecasting models. Some comparisons, like those between model specifications with the same dependent variable, are invariant to the choice of benchmark model used to rescale the root mean square errors.

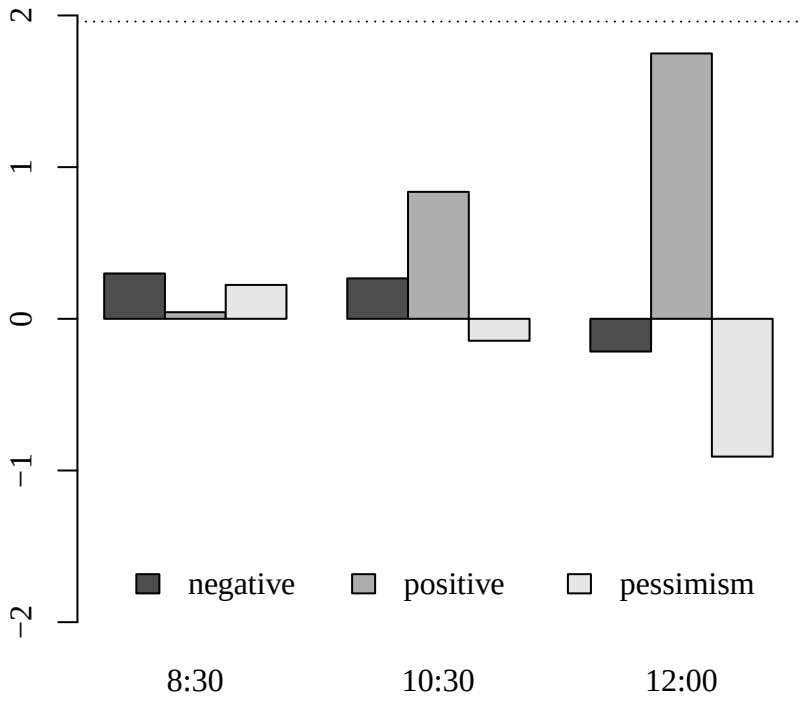

Figure 1: White [1980] t-statistics associated with ordinary least squares estimates of the coefficient β in regressions of type shown in equation (4) and estimated in Garcia [2013]. The horizontal axis indicates the starting time that I use to compute post-open returns. I compute post-open returns until 5:30 p.m.. A dotted horizontal line appears at 1.96, the upper 97.5^{th} percentile of the standard normal distribution.

and Garcia [2013], the media factor f_t is the number of occurrences of negative or positive words, or the difference between the numbers of occurrences of negative and positive words ('pessimism'), in newspaper articles matched to stock returns on day $t+1$.[16]

I compute t-statistics using White [1980] standard errors and present the t-statistics associated with the coefficient β in Figure 1. The t-statistic with the largest absolute value is 1.74 and occurs when counts of positive words are used to explain the next day's change in stock index futures price between 12:00 p.m. and 5:30 p.m.. As Figure 1 makes clear, the relationship between the media factor and stock returns found in Tetlock [2007] and Garcia [2013] is not replicated on these data, for two reasons. First, the signs of the estimated β coefficient from equation (4) do not always agree with their hypothesised sign. The number of negative words is positively associated with the next day's 8:30 and 10:30 a.m. post-open stock return, and the 'pessimism' factor, calculated as the difference between the numbers of negative and positive words, is positively associated with the next day's 8:30 a.m. post-open stock returns. The number of positive words is negatively associated with the next day's 10:30 a.m. post-open stock return. Second, not one of the β coefficients from equation (4) are estimated to be statistically different from zero at the 5% significance level.[17] If a lack of significance were due to fewer observations in my sample,[18] then I would expect to see similar estimates of the coefficient β in my sample to those obtained in the earlier studies, while obtaining larger standard errors. However, I obtain estimates of β of between -2.1 and 4.1 basis points per unit of standard deviation of f_t, which is about half the size of the estimates obtained in these earlier studies.

The fact that the results of Tetlock [2007] and Garcia [2013] are not replicated on my data could reflect problems with their estimation like finite sample bias or reverse causality, or that fact that the single print media factor model is a purely historical phenomenon, as explained in Section 1. Alternatively, the results that they find could be specific to the newspapers that they study, which is suggested by the lack of predictability found for internet stock message boards and company press releases [Antweiler and Frank, 2004, Luss and d'Aspremont, 2015], or specific to the large, relatively closed economy that is the United States.

A final explanation for the lack of explanatory power of the single media factor on this sample is that specification (4) omits other important media factors. The effect that earlier studies attribute to the single media factor could then reflect to some extent the role for such omitted factors. To explore such an explanation, I present results from estimating (4) with a generalised representation of the print media factor f_t that nests the single print media factor as a special case.

4.2 Multiple media factors

4.2.1 Informativeness for post-open returns

The poor performance of the single media factor specification (4) on these data motivates the consideration of media factors that could be omitted from that specification. For example, rather than counting the number of occurrences of all positive words on a given day, we could count the number of occurrences on that day of each positive word in a predefined list of positive words. By considering a vocabulary made up of positive and negative words, I obtain a richer set of explanatory variables that nest the single factor time series explanatory variable in specification (4). Rather than restricting attention to whether a single factor constructed from news can price past levels of the stock market, I consider which groups of words can price future levels of the stock market before having observed them.

[16] The matching procedure is explained in Section 3 and ensures that articles matched to returns on trading day t are published on calendar day $t-1$ or earlier. I use positive and negative words as defined in Loughran and McDonald [2011].

[17] Starting times other than the three presented in Figure 1 and ending times other than 5:30 p.m. produce t-statistics that are closer to zero in absolute value and similarly incorrectly signed.

[18] I have a maximum of 759 observations in these regressions, while Tetlock [2007] and Garcia [2013] have $3,709$ and $19,184$ observations respectively.

Table 1: Minimum relative cumulative root mean square (out-of-sample) prediction errors for intraday returns on Top 40 index futures contracts. The rows of the table index different models, in the sense of different sets of explanatory variables used for prediction. All models with lagged returns also include lagged squared returns and day-of-the-week explanatory variables, as discussed in the text. The columns of the table indicate whether intraday returns being predicted are computed from 8:30 a.m., 10:30 a.m. or 12:00 p.m., while all intraday return windows end at 5:30 p.m.. Each entry in the table gives the minimum, across the various newspaper vocabularies described in the text and tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2$, of the root mean square prediction error, expressed relative to the root mean square prediction error of a mean-only model.

model	8:30	10:30	12:00
best model using news only	0.91	0.98	1.04
best model using returns only	1.01	1.03	0.98
best model using returns and news	1.01	1.02	0.96

Table 1 presents the minimum, across the six newspaper vocabularies and tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2$, of the full-sample cumulative root mean square prediction error, given in display (3) with $l' = 100$, expressed as a ratio to the full-sample cumulative root mean square prediction error of a benchmark model that forecasts the historical mean return, for predicting changes in the natural logarithm of Top 40 futures prices over an interval starting at 8:30 a.m., 10:30 a.m. or 12:00 p.m. and ending at 5:30 p.m.. The best performing model in the table predicts the intraday return between 8:30 a.m. and 5:30 p.m. and achieves a cumulative root mean square prediction error that is 91% of that of the model that forecasts using the historical mean return. This relative root mean square prediction error can be interpreted as 9% lower than the variation of the post-open stock returns being predicted, or equivalently as 9% lower than the root mean square prediction error of a naive model that forecasts using the historical mean return. The vocabulary for this best-performing model contains all words appearing in the newspaper archive and the model excludes lagged return features (i.e. it excludes lagged returns, lagged squared returns and day-of-the-week indicator variables). Note that this vocabulary outperforms vocabularies constructed from pairs and triples of words and those vocabularies constructed from the Loughran and McDonald [2011] dictionary. This suggests that the predictive content of news for future returns can be distilled into individual words, rather than pairs or triples of words, but cannot be reduced simply to those individual words with positive or negative connotations.[19] By reading across the first row of Table 1, we see that predictability of the previous day's news for intraday returns declines as the trading day progresses, so that by noon the previous day's news is incorporated into the Top 40 futures price.

The results in the first row of Table 1 demonstrate the predictive content of news for 8:30 and 10:30 a.m. post-open stock returns. In addition, we may be interested to know whether this predictability can be explained by the predictive content of lagged returns, volatility or day-of-the-week effects. In particular, the content of news articles could reflect past stock returns and past volatility, which could be related to future stock returns. If we were to find that news features provide predictive content by proxying for other features of stock returns, this would not change the predictive content of news features, but could offer an explanation of the mechanism by which news is informative.

The second row of Table 1 shows the minimum, across the tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2$, of the full-sample cumulative root mean square prediction error of a model based on lagged returns, volatility and day-of-the-week effects only relative to that of a model that forecasts the historical mean return, for each post-open return window. The last row of Table 1 also presents such a relative prediction error, but takes the minimum over the six newspaper vocabularies and four choices for the

[19]Note that there is no concern of reverse causality in the relationship between news and returns described here because only news from the previous calendar day or earlier is matched to the intraday return on a given day. As described in Section 2, the opening price at 8:30 a.m. is determined by an opening auction and differs from the previous day's close price. Hence it is unlikely that any news published after the previous trading day's 5:30 p.m. market close would not already be incorporated into the 8:30 a.m. post-auction opening price.

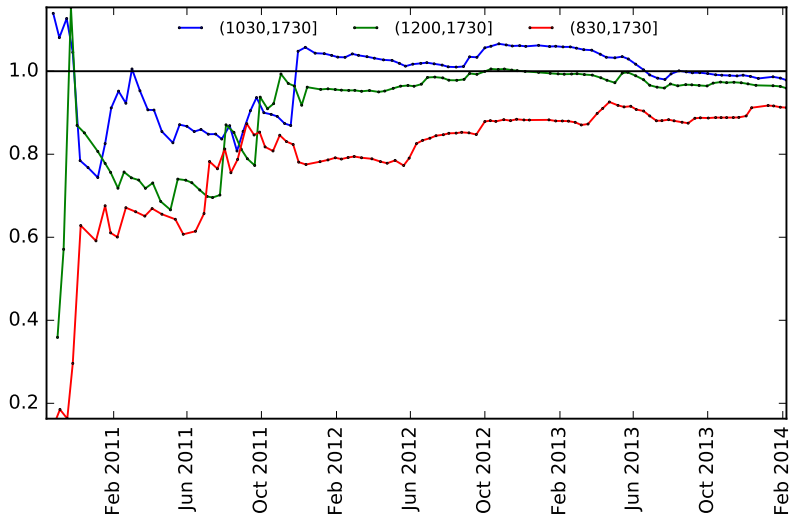

Figure 2: Relative cumulative root mean square prediction error based the best performing model specifications in Table 1. The figure plots the expression in display (3) as a ratio of the cumulative root mean square prediction error of a model that forecasts the historical mean return, against l'. The figure shows these prediction errors for post-open returns between 8:30 a.m. and 5:30 p.m., between 10:30 a.m. and 5:30 p.m. and between 12:00 p.m. and 5:30 p.m..

tuning parameter C and applies to a model containing all of the above explanatory variables. By comparing the entries in the last two rows in any given column, we see that the inclusion of news does not worsen the fraction of the out-of-sample standard deviation of post-open returns that can be explained by a model containing lagged returns, volatility and day-of-the-week explanatory variables. This confirms that the above predictability of news for stock returns does not appear to proxy for known sources of predictability like lagged returns, volatility or day-of-the-week effects.[20]

A further observation from Table 1 is that the best performing model for each type of post-open return always involves news features. This suggests that no matter the definition of post-open returns, it is always advantageous to have news for prediction. More specifically, the relative prediction error for 8:30 a.m. post-open stock returns using returns only is larger than the that of the best performing model involving news features by an amount equal to 10 percent of the variation of such stock returns.[21] For 10:30 a.m. post-open returns, this figure falls to 5 percent, and for 12:00 p.m. post-open returns it falls again to 2 percent. In this sense, the informativeness of news features for intraday stock returns declines with time.[22]

The preceding discussion concerned full-sample root mean square prediction errors, i.e. those in display (3) with $l' = 100$, but this does not give an indication of the stability or instability of such prediction errors. Figure 2 plots the relative cumulative root mean square prediction errors for two types of post-open returns against the time index $l' = 1, 2, \ldots, 100$. At each date on the horizontal axis, the height of a line in Figure 2 shows the (relative) cumulative out-of-sample root mean square prediction error up to that date, based on the errors from fitting models over expanding windows and predicting out of sample one day ahead. The right endpoint of the lines in Figure 2 correspond to the best full-sample relative cumulative root mean square prediction errors presented in Table 1, and

[20]Note that the out-of-sample predictive performance of a model does not necessarily improve with the addition of more explanatory variables. This behaviour is similar to that of the well known adjusted R^2, which can decline with the addition of more explanatory variables.

[21]That is, $10\% = 1.01 - 0.91$ in the first column of Table 1.

[22]The calculations are $5\% = 1.01 - 0.98$ and $2\% = 0.98 - 0.96$ respectively.

Table 2: Informativeness of the multiple media factors for daily stock returns, stock trading volume, bond returns, the equity premium, the term premium and capital flows n steps ahead for $n = 1, 2, 3, 4, 5$. Each entry in the table represents the difference between one and the minimum, over the six vocabularies and four tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2$, of the ratio of the root mean square (out-of-sample) prediction error of the news-only prediction model to the root mean square (out-of-sample) prediction error of a historical mean forecast model. The entries are multiplied by 100 to be expressed in percentage point units. Entries that would be negative are not shown with the understanding that the news-only model is not more informative than the historical mean model in these cases. The calculation of each variable is described in Section 2.

	n				
Variable	1	2	3	4	5
Top 40 return	12	15	16	7	·
Top 40 change in log volume	56	54	54	55	44
GOVI return	5	·	3	·	6
Equity premium	13	22	17	7	·
Term premium	6	·	9	9	8
Debt net inflow	·	1	11	·	·
Equity net inflow	6	24	8	16	10
Debt and equity net inflow	·	2	12	·	3

the trajectory of the lines shows more detail about the stability of these full-sample estimates. The estimates of relative cumulative root mean square prediction error for post-open stock returns stabilise from about February 2012. This stability suggests that these errors are relatively precisely estimated and would not increase dramatically with the addition of more data. Therefore, the predictive content of news articles for post-open stock returns does not seem to be explained by overfitting in-sample or out-of-sample.

Among the collection of news-only models in the first row of Table 1, the vocabulary that produces the best relative cumulative root mean square prediction error is the list of all individual words. For the combined model of news and lagged returns, the vocabulary based on all individual words and pairs of consecutive words performs equally as well as the vocabulary based on all individual words, pairs of consecutive words and triples of consecutive words. These vocabularies perform best for of the choices of post-open return in Table 1. In no case do the three vocabularies of Loughran and McDonald [2011] emotive words outperform the three vocabularies of all individual, pairs and triples of words.[23] Therefore, the informativeness of news for post-open stock returns seems to come from the non-emotive words. This finding suggests that the single media factor model in specification (4) performs poorly due to an omitted factor.

4.2.2 Informativeness for daily returns, trading volume and capital flows

The preceding discussion concerned the informativeness of print news media for post-open stock returns. I now turn to the informativeness of print news media for daily stock and bond returns, change in trading volume, the equity premium, the term premium and capital flows. I investigate the informativeness of print news media for these variables up to 5 trading days ahead.

Table 2 shows the reduction in variance of each variable at each forecast horizon that is achievable using only the multi-factor representation of print news media. The entry in this table for a given variable and forecast horizon $n = 1, 2, 3, 4, 5$, shows the maximum, over the six vocabularies and four tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2$, of the percentage reduction in daily variance of the variable that

[23] In the case of 8:30 a.m. post-open returns and news-only models, models based on all six vocabularies, except the negative words only and positive words only vocabularies, achieve the same relative cumulative root mean square prediction error of 91%. However, the performance of the combined positive and negative words vocabulary is sensitive to the choice of the tuning parameter C while the other vocabularies are not.

is achievable using only the multi-factor representation of the print news media n days earlier. The overall impression from the table is that the print news media is informative about a number of these variables. I discuss this table row-by-row below, including how the results change with the addition of the extra explanatory variables specified in Section 3.2.

The first row of Table 2 shows that the multi-factor representation of print news media is informative for daily stock returns up to four days ahead, but not five days ahead. The multliple media factors reduce the one-day ahead out-of-sample variation of daily stock returns by 12% of their original level, which is slightly larger than the 9% achieved above for post-open returns, suggesting that the difference is due to reverse causality between out-of-hours financial market developments and print media content. Not shown in the table, a large portion of this reduction in daily variation does not seem to be explained by serial correlation, volatility or day-of-the-week effects in the sense that the best model with only serial correlation, volatility and day-of-the-week explanatory variables only achieves a 3% reduction in daily variation. For two-, three- and four-day ahead returns, none of the 15%, 16% and 7% reduction in daily variation shown in Table 2 seems to be attributable to serial correlation, volatility or day-of-the-week effects in the sense that the best models in each case with only serial correlation, volatility and day-of-the-week effects cannot explain any of the daily variation in returns.[24]

The vocabularies that achieve the best performance shown in first row of Table 2 also vary with the forecast horizon n. The most informative vocabulary for one-day ahead returns, which achieves the 12% reduction in daily variance shown in the first row and column of Table 2, is the list of negative words. For two-day ahead returns, there is not much difference between the six vocabularies, with four achieving the same 15% reduction in daily variation shown in Table 2. However, for three- and four-day ahead returns, the best performing vocabularies are the more complex non-emotive vocabularies. The vocabulary of all individual words and pairs of consecutive words and the vocabulary of all individual words, pairs of consecutive words and triples of consecutive words both achieve the 16% reduction in daily variance for three-day ahead returns in Table 2. The most informative vocabulary for four-day ahead returns is the vocabulary of all individual words, pairs of consecutive words and triples of consecutive words.

It therefore seems that the print news media plays an important role in the pricing of stocks. It takes about four days for the effects of the print news media to be incorporated into stock prices and the effects are clearest three days ahead. Emotive words are the most important characteristics of print news media in affecting stock prices one day ahead, while more complex non-emotive features play a bigger role three and four days ahead.

Closed economy models of heterogeneous agents predict that trading volume is determined by the extent of disagreement between these agents. The second row of Table 2 shows that the multi-factor representation of print news media can explain 56% of the out-of-sample variation in trading volume changes one day ahead, and this fraction declines to 44% of the variation five days ahead. These fractions are large, but a large proportion could be attributed to the variation explicable by lagged stock returns, volatility day-of-the-week effects. Not shown in the table, the best models with lagged stock returns, volatility and day-of-the-week effects explain 47% and 50% of the variation in trading volume changes one and two days ahead respectively, leaving only the remaining 9% and 4% of the variation in trading volume changes explicable by print news media only. A similar calculation reveals no substantive extra explanatory power three, four or five days ahead.[25] Therefore, I find a role for print news media in affecting trading volume up to two days ahead, which is consistent with a role for print news media in heterogeneous agent models, but I cannot find a role for print news media in affecting trading volume three days ahead where the affect of print news media on stock prices is strongest.

The multi-factor representation of print news media is even more informative about daily variation in trading volume changes for stocks in the All-Share Index. However, the extra explanatory power of print news media over the lagged return, volatility and day-of-the-week explanatory variables is

[24] I obtain the same results using daily returns on the All-Share Index.

[25] The exact fractions of (out-of-sample) variance explicable by the best models with only lagged stock returns, volatility and day-of-the-week explanatory variables are 56%, 53% and 44% for three-, four- and five-day ahead returns respectively.

smaller in the case of the All-Share Index. This finding suggests that the role for print news media in driving trading in heterogeneous agent models is primarily through large stocks. This finding favours an interpretation where the print news media affects some agents' beliefs about large stocks, rather than affecting some agents' risk aversion toward all risky assets.

The third row of Table 2 shows limited evidence of the informativeness of print news media for government bond prices. By contrast, the fourth row of this table shows that the multi-factor representation of print news media is appreciably informative for the equity premium, with the greatest reduction in (out-of-sample) variance being achieved two days ahead.[26] These results suggest that the print news media is primarily informative for the prices of domestic risky assets rather than domestic risk-free assets.

If we regard long-term bonds as the risky asset and short-term bonds as the risk-free asset, we may expect the print news media to affect the relative pricing of such bonds for the same reasons as we would expect the print news media to affect the pricing of stocks relative to bonds above. The fifth row of Table 2 shows that the multi-factor representation of print news media is informative about (out-of-sample) variation in the excess return of long-term bonds over short-term bonds one, three, four and five days ahead. After allowing for serial correlation, volatility and day-of-the-week effects, 7%, 3% and 2% of the three-, four- and five-day ahead predictability remain, and none of the one-day ahead predictability remains. Hence, there appears to be a role for print news media in affecting the price of long-term bonds relative to that of short-term bonds, and this effect is strongest three days ahead.

The last three rows of Table 2 show that the multi-factor representation of print news media is informative about net portfolio capital flows into South Africa. This informativeness is strongest for equity flows two days ahead and debt flows three days ahead. The informativeness of print news media for total debt and equity net portfolio inflows mirrors such informativeness for debt flows because debt flows tend to be several multiples larger than equity flows.

The fraction of the (out-of-sample) variance of two-day ahead equity portfolio flows that can be explained by the multi-factor representation of print news media but not by lagged stock returns, volatility or day-of-the-week variables is 4%. The vocabularies of positive words and combined positive and negative words are the most informative for equity portfolio flows two days ahead. These findings should be compared with the informativeness of print news media for two-day ahead stock returns and two-day ahead stock trading volume, which are also driven by emotive vocabularies. These findings are consistent with foreign residents affecting domestic stock prices by adjusting their demand for domestic stocks on any given day based on the emotive content of domestic print media two days earlier.

The best model for three-day ahead debt portfolio flows using only lagged bond returns, volatility or day-of-the-week variables cannot explain any of the variation in this dependent variable. Therefore, the 11% of the variance of three-day ahead bond portfolio flows explicable by the multi-factor representation of print news media does not seem to reflect lagged bond returns, volatility or day-of-the-week effects. This explanatory power also coincides with the informativeness of print news media for the three-day ahead term premium, which suggests that foreign residents' demand for long-term relative to short-term domestic bonds on any given day depends on the content of domestic print media three days earlier.[27]

5 Conclusion and proposed future work

The print news media could play a role in asset pricing by affecting agents' beliefs or preferences. These affected beliefs and preferences could be of all agents or of only some agents, and these agents in

[26] None of the (out-of-sample) variation in the equity premium is explained by serial correlation, volatility or day-of-the-week effects.

[27] No particular vocabulary seems to drive the predictive content of the multi-factor representation of domestic print media for three-day ahead portfolio debt flows.

turn could be foreign or domestic.[28] Previous work has explained the role for print news media in the pricing of stocks in terms of emotive language affecting either the beliefs of some irrational domestic agents, or the preferences of some susceptible domestic agents, in a large closed economy. I document some uncertainties with this conclusion owing to potential finite sample biases, reverse causality or the discovery of a purely historical phenomenon.

In this paper, I use data on a small open economy to show evidence that confirms the importance of print news media for the pricing of stocks and bonds. Specifically, I find that a multi-factor representation of print news media can predict about 9 percent of the variation in daily stock returns (one day ahead) and 7 percent of the variation in the daily term premium (three days ahead), but only limited quantities of the variation in daily aggregate bond returns. This role for the print news media does not seem to reflect a purely historical phenomenon, finite-sample biases, reverse causality, serial correlation, volatility or day-of-the-week effects, which therefore rules out the sources of uncertainty associated with previous work.

I also present some evidence of three mechanisms by which these overall effects could operate. First, the excess predictability of print news media for stock returns up to two days ahead is driven by emotive language and large stocks, and it coincides with the excess predictability of print news media for equity capital inflows and trading volume up to two days ahead. This finding supports an interpretation of heterogeneous agents that trade with each other on the basis of recent (i.e. two days' prior) emotive language in the print news media. The finding also suggests that some of the agents are located abroad, which is consistent with an information asymmetry between foreign and domestic residents about domestic stocks. The clearer effect for larger stocks suggests either that emotive language in the domestic print news media influences foreign residents' beliefs about large stocks in particular, or that foreign residents act on their beliefs about all domestic stocks through those that have lower transaction costs, but is more difficult to reconcile with interpretations of the print news media affecting foreign residents' risk aversion because risk aversion would affect both large and small stocks.

Second, the excess predictability of print news media for stock returns three and four days ahead is driven by non-emotive language and large stocks, and does not coincide with any excess predictability of print news media for equity capital inflows or trading volume at those forecast horizons. This finding supports an interpretation where agents' beliefs and preferences respond homogeneously to non-emotive language in the print news media three to four days prior. While the preceding finding suggests a short-term role for heterogeneous 'sentiment' between domestic and foreign residents, this finding suggests that other 'non-sentiment' factors associated with the print news media drive its medium-term effects on stock returns.

Third, the excess predictability of print news media for the term premium three days ahead is driven by non-emotive language and coincides with excess predictability of the print news media for portfolio debt net inflows from abroad. This finding supports an interpretation where foreign agents' demand for domestic long-term bonds depends on non-emotive language three days prior, where foreign agents finance their purchases of domestic long-term bonds through selling domestic short-term bonds, and where foreign agents invest their proceeds from selling domestic long-term bonds in domestic short-term bonds.

References

Werner Antweiler and Murray Z Frank. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.

John Y Campbell, Sanford J Grossman, and Jiang Wang. Trading volume and serial correlation in stock returns. *Quarterly Journal of Economics*, 108(4), 1993.

[28] A role in asset pricing for heterogeneous agents should be understood to be accompanied by constraints on the agents whose beliefs or preferences are not subject to outside influences.

David Cutler, James Poterba, and Lawrence Summers. What moves stock prices? Technical report, National Bureau of Economic Research, 1988.

Zhi Da, Joseph Engelberg, and Pengjie Gao. The sum of all FEARS: Investor sentiment and asset prices. *Review of Financial Studies*, 28(1):1–32, 2015.

J Bradford De Long, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann. Noise trader risk in financial markets. *Journal of Political Economy*, pages 703–738, 1990.

Marcel Fratzscher. Capital flows, push versus pull factors and the global financial crisis. *Journal of International Economics*, 88(2):341–356, 2012.

Diego Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics, 2009.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

Ronny Luss and Alexandre d'Aspremont. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6):999–1012, 2015.

Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.