Erasmus School of Economics

BSc^2 Economics/Econometrics Thesis

The Impact of News on Returns Worldwide

J.J. Bal (426991)

Under supervision of X. Xiao

Second Assessor: X. Gong

Abstract

This paper examines the impact of sentiment and frequency of news articles on monthly and 12-month stock returns in 51 countries. I show that the results of Calomiris and Mamaysky (2019), who found significant predictive value of news variables, hold when corrected for severe overlapping data issues present in their research. However, large variation in predictive value is revealed by country-level Ordinary Least Squares (OLS), Least Angle Regressions (LARS), and Weighted Least Squares (WLS) procedures. Furthermore, I find that United States' data is on average better at predicting a country's stock returns than data of the country itself.

July 7, 2019

Contents

1 Introduction

The impact of information on stock market returns has been investigated for decades. In the earlier years, research did not find much evidence that returns could be predicted using available information (for instance, Fama, Jensen, & Roll, 1969). However, as time progressed, an increasing number of instances were brought to light that challenged this view. In a review published in 1978, Jensen discusses the research revealing some of these instances, such as Charest (1978) who shows that the announcement of dividend changes can predict U.S. stock market returns for several months. Even the founding father of the Emerging Market Hypothesis and Nobel laureate Eugene Fama finds that stock returns can partially be predicted using information on expected inflation (Fama and Schwert, 1977).

In recent years, the question whether stock returns can be predicted is still not entirely settled (Spiegel, 2008). Welch and Goyal (2008) find that economic variables fail to deliver forecasting gains, whereas Rapach, Strauss, & Zhou (2010) claim to have found predictive ability of 15 economic variables. The disagreement and contrary findings on a topic as important as forecasting stock returns make further research highly relevant and hence this is conducted by many researchers.

In a very interesting attempt to forecast returns, a group of researchers has focused on investigating the predictive value of news on aggregate. What makes these techniques unique is rather than investigating how one news event is digested by the market (for example Fama et al., 1969; Charest, 1978), they examine the impact that trends in all news articles combined have on stock returns. One paper that falls into this category is the research by Calomiris and Mamayksy (2019), who predict returns in stock markets in 51 countries, split in developed markets (DM) and emerging markets (EM) groups. They find that self-obtained metrics on news such as topic-specific sentiment, frequency and unusualness of word flow predict stock returns for one year ahead, as measured by *R-squared*, the increment in *R-squared* when news variables are added to a baseline regression, and individual significance of the coefficients of part of the news variables. This finding, if econometrically justifiable, has large implications for the financial literature, as it shows that news has additional forecasting value over the economic variables used by Rapach et al. (2010).

However, the paper raises some questions and leaves others unanswered. First of all, the paper fails to properly address the issue of overlapping data. The authors construct the 12-month ahead returns by compounding 12 consecutive monthly returns. The next data point is constructed by dropping the first month and adding a month at the end. This causes two consecutive 12-month returns to have eleven underlying monthly returns in common, creating

overlapping returns. The overlapping returns are problematic from an econometric point of view, as they generate a moving average (MA) error term that causes Ordinary Least Squares (OLS) parameter estimates to be inefficient and hypothesis tests to be biased (Hansen and Hodrick, 1980). Hence, the results obtained by Calomiris and Mamaysky (2019) are not reliable and need to be corrected to account for overlapping returns. This paper does exactly that by employing a Weighted Least Squares (WLS) procedure developed by Johnson (2018) that corrects for overlapping returns. It is of utmost importance that the results of Calomiris and Mamaysky (2019) are reevaluated while correcting for overlapping returns, as it could change the conclusions that can be drawn from the research and therefore significantly impact academic literature.

Although Calomiris and Mamaysky (2019) claim to have found significant forecasting value of news variables for emerging market countries and developed market countries as a group using panel regressions, they do not give insight into any country-specific differences. In order to provide this insight, this paper will perform regressions on country level. The first regressions that will be performed are country-level OLS regressions, which can be seen as the individual equivalent of panel regressions. Subsequently, the importance of news variables in each country will be assessed by restricting the sum of the absolute values of the coefficients in the OLS regression using a Least Angle Regression (LARS) algorithm, which is similar to a "lasso" procedure (Efron, Hastie, Johnstone, & Tibshirani, 2004; Tibshirani, 1994). Finally, the WLS regressions that correct for overlapping returns will be performed.

Calomiris and Mamaysky (2019) only report the regression results for 12-month ahead returns and omit the results for monthly returns. The motivation for this choice is that they have found little forecasting power for monthly returns compared to 12-month returns. However, the relatively superior performance of the 12-month returns could be due to the overlapping data that is present in this sequence. In order to see how the prediction of 12-month returns corrected for overlapping data issues compares to monthly returns, this paper will perform the panel regressions, OLS regressions and LARS regressions for monthly returns as well.

The above can be captured with the following research questions and subquestions:

1. *Are the news variables of Calomiris and Mamaysky (2019) still of predictive value when corrected for overlapping data issues?*

    (a) *Is the value of R-squared and the increment in R-squared when including news variables of similar magnitude in WLS regressions compared to the panel regressions of Calomiris and Mamaysky (2019)?*

    (b) *Do the coefficients of individual news variables remain significant in the WLS regres-*

sions?

2. *Are there large differences in forecasting value of news variables on a country level as indicated by OLS, LARS and WLS regressions?*

In addition, a third research question will be assessed that focuses on the impact of the United States on other stock markets worldwide. Calomiris and Mamaysky (2019) use millions of news articles in all 51 countries to construct news variables for each country, but this procedure may be unnecessarily complex. Rapach, Strauss, & Zhou (2013) found that lagged U.S. returns were able to significantly predict returns in other industrialized countries. For reasons explained in section 2, stock market returns in the United States seem to have a leading role compared to other stock markets. It is interesting to see whether U.S. news variables possess a similar leading role and are able to predict stock returns in other countries. If this is the case, only U.S. news data needs to be collected to predict worldwide returns, which is far more efficient. I formulate the following research question to assess this:

3. *To what extent are the United States' news variables able to predict returns in other countries?*

The outline of this paper is as follows. I begin by discussing existing literature on return forecasting using news metrics in section 2. Section 2 also examines the (leading) role of the U.S. and its stock market in the world. Next, section 3 discusses the data used in the research, which corresponds to the data of Calomiris and Mamaysky (2019). Section 4 elaborates on the methodology, first discussing the methodology to replicate the results of Calomiris and Mamaysky (2019) and subsequently discussing the methodology to perform the OLS, LARS and WLS regressions. Section 5 contains the results of the regressions, which will be used to answer the research questions in the concluding section 6.

2 Literature Review

As mentioned in the introduction, Calomiris and Mamayksy (2019) attempt to extract summary statistics from a large database of news articles and subsequently use these statistics to predict future stock returns.

The first paper that established a link between news content and stock returns was Tetlock (2007). In this paper, Tetlock shows that a measure of pessimism in a popular Wall Street Journal column predicts downward pressure on U.S. stock prices the next day. However, he also shows that this initial negative effect is offset in the following four trading days, rendering the

impact on weekly returns to be insignificantly different from zero. Tetlock, Saar-Tsechansky, & Macskassy (2008) expand on this by showing that a measure for the number of negative words on a day in several large newspapers can predict Dow Jones Industrial Average (DJIA) returns of the following day.[1] Later work confirms the findings of Tetlock, with Garcia (2013) being a recent example.

Interestingly, all literature discussed so far has focused on short-term predictions, i.e. next day returns. Here, Calomiris and Mamaysky (2019) differentiate themselves as they predict monthly and yearly returns. Since, Tetlock (2007) and later Garcia (2013) have shown that the impact of daily news on stock returns is reversed within a week, Calomiris and Mamayksy (2019) accumulate news monthly in order to boost its long-term predictive value. They find support for this attempt from Sinha (2016) and Heston and Sinha (2017), who have shown that weekly news predicts returns for 13 weeks. Given this result, accumulating news monthly in order to predict monthly and yearly returns makes sense as a next step in the field. Furthermore, Calomiris and Mamayksy (2019) extent previous literature by including 51 countries in their analysis, where previously only the United States was considered. They split this sample of 51 countries in a developed markets (DM) group and an emerging markets (EM) group. All results are obtained for each group separately. This division is made in order to account for the fact that returns are influenced by different factors in DM and EM countries. Hence, it can be expected that news on a particular topic has a different impact on stock returns in DM countries compared to EM countries. These differences in impact are caused by differences in underlying circumstances, such as political climate, market liquidity, and legal environment (Calomiris, Love, & Pería, 2012; La Porta, Lopez-de-Silanes, Shleifer, & Vishny, 1998).

Lastly, literature will be discussed that has inspired the third research question. As mentioned in the introduction, Rapach et al. (2013) discovered that lagged U.S. stock returns are significant predictors of stock returns in other developed countries. According to the authors, Rizova (2010) provides the most probable economic explanation for this finding. Suppose country A and B are major trading partners. Rizova (2010) finds that stock market returns of country A predict exports to country A from country B as well as imports from country A by country B. However, the implications of the stock market returns of country A on exports and imports are not immediately digested by the stock market of country B. As a result, stock returns in country A are able to forecast future stock returns in country B. Because the U.S. is a major trading partner for many developed countries, this theory explains the findings of Rapach et al. (2013). The theory is consistent with gradual information diffusion, which states that information flows

[1]The authors define the following variable: $Neg = \frac{\text{No. of negative words}}{\text{No. of total words}}$. Subsequently, the daily return of the Dow Jones Index of the next day is regressed on the standardized value of *Neg*.

slowly from assets that receive a lot of investor attention (e.g. U.S. stocks) to related assets, which leads to return predictability.

If the hypothesis of gradual information diffusion holds, one may expect that U.S. news has a strong predictive value for future stock returns in other countries as well. Hence, rather than obtaining news variables for each country separately, it may suffice to collect U.S. news to predict monthly and 12-month returns for all countries. As this significantly decreases the amount of information that needs to be obtained, it is undoubtedly worth investigating. It is important to note that Rapach et al. (2013) only focused on industrialized (developed) countries, leaving emerging countries out of the picture. As emerging countries on average trade less with the U.S. than the developed countries, the forecasting power of U.S. news variables may be weaker.

3 Data

Calomiris and Mamaysky (2019) investigate the impact of news on returns for 51 countries in the period April 1998 - December 2015. The countries are split into two groups: emerging market countries and developed market countries, which can be found in Table 1. For each of these countries, three types of data are obtained: return, macro and news data.

The return data (*return*) used as the dependent variable in the regressions are monthly total returns of a country's major stock index. One-year returns ($returns^{12}$) are obtained by accumulating the next twelve monthly returns. Hence, two subsequent one-year returns have eleven underlying monthly returns in common, causing overlapping data issues. Monthly volatility (*sigma*) of the returns is approximated by the realized volatility as reported by Bloomberg over the last 20 trading days of the month. As a proxy for *value* in month t, the average level of the country's stock market index from 5.5 years to 4.5 years before month t is divided by month t's closing price. This method is preferred over methods that use accounting measures such as the book value of equity because the latter would significantly limit the number of observations caused by a lack of available data.

Table 1: Countries and country codes
List of countries in developed market (DM) countries and emerging market (EM) countries, together with their Thomson Reuters codes (TR codes).

	Emerging markets			Developed markets	
	Country	TR code		Country	TR code
1	Argentina	AR	1	Australia	AU
2	Brazil	BR	2	Austria	AT
3	Chile	CL	3	Belgium	BE
4	China (PRC)	CN	4	Canada	CA
5	Colombia	CO	5	Denmark	DK
6	Czech Republic	CZ	6	Finland	FI
7	Estonia	EE	7	France	FR
8	Ghana	GH	8	Germany	DE
9	Hong Kong	HK	9	Greece	GR
10	Hungary	HU	10	Ireland	IE
11	India	IN	11	Italy	IT
12	Indonesia	ID	12	Japan	JP
13	Israel	IL	13	Luxembourg	LU
14	Kenya	KE	14	Netherlands	NL
15	Malaysia	MY	15	New Zealand	NZ
16	Mexico	MX	16	Norway	NO
17	Nigeria	NG	17	Portugal	PT
18	Peru	PE	18	Singapore	SG
19	Philippines	PH	19	Spain	ES
20	Poland	PL	20	Sweden	SE
21	Russia	RU	21	Switzerland	CH
22	Slovakia	SK	22	United Kingdom	GB
23	Slovenia	SQ	23	United States	US
24	South Africa	ZA			
25	South Korea	KR			
26	Thailand	TH			
27	Turkey	TR			
28	Ukraine	UA			

In order to prevent an omitted variable bias and isolate the direct impact that news has on returns, a large list of macro variables is used as regressors. The macro data have been obtained from numerous sources, including the International Monetary Fund (IMF) and the World Bank. The following macro variables have been used in the analysis: year-over-year GDP growth (*gdp*), year-over-year inflation (*gdpdeflator*), private sector credit to GDP (*cp*), year-over-year change in cp (*dcp*), local currency interest rate (*rate*), percent US$ appreciation against local currency (*dexch*), and pre- and postelection dummies (*pre* and *post*). However, given that pre- and postelection dummies are rarely significant in the regressions of Calomiris and Mamaysky (2019) and very time consuming to obtain, I will not include them in my analyses. Table 2 provides an overview of all variables together with their descriptions, abbreviations, and data sources.

Table 2: Descriptions of variables and article topics
The upper table contains descriptions of the variables used in the regressions and the corresponding sources. Below the article topics are described. The *Macro* topic only applies to Emerging Market (EM) countries and *Credit* only to Developed Market (DM) countries.

Variable	Description	Source
return	Monthly stock return (in %) including capital gains and dividend	Bloomberg
return^{12}	One-year ahead stock returns	Bloomberg
sigma	Rolling 20-day realized volatility reported in annualized terms	Bloomberg
value	Stock index level from 4.5 to 5.5 years ago divided by current stock index level	Bloomberg
gdp	Real GDP growth rate	IMF
gdpdeflator	Rate of change of the GDP deflator	IMF
cp	Private sector credit-to-GDP ratio	World Bank
rate	Local currency interest rate: monthly deposit rates are used for EM and ten-year government bond yields for DM	IMF, Datastream
dexch	Percentage change in value of US Dollar compared to the local currency	Bloomberg, Datastream
entropy	Daily measure of the unusualness of words used, averaged over a month	Website of H. Mamaysky
artcount	Number of articles written about a country per dag, averaged over a month	Website of H. Mamaysky
s[Topic]	Topic-specific sentiment of articles in a given month	Website of H. Mamaysky
f[Topic]	Topic-specific frequency of articles in a given month	Website of H. Mamaysky

Topic	Description
Mkt	Markets-related articles
Govt	Government-related articles
Corp	Corporate governance and structure related articles
Comms	Commodities-related articles
Macro	Macroeconomic articles
Credit	Articles regarding the extension of credit

The authors use multiple variables to capture the potential explanatory value of news, all calculated for each country individually. The first variable is *artcount*, which is simply the number of articles per country per month. In order to capture possible information inherent to the unusualness of an article, the authors define a variable named *entropy*. The computation of this variable is fairly complex, but the intuition behind it is that the value will increase when articles in a certain month contain a lot of unusual language and word combinations that have not been witnessed in the past. The authors reason that unprecedented language could be used to describe new market or economic phenomena, which could impact returns or risk in the future, hence giving *entropy* explanatory value.

A question that the authors aim to answer in the paper is whether the topical context of an article matters for the impact of the news. In order to answer this question, five topics have been defined for emerging markets and developed markets separately using a word cluster algorithm that follows the approach of Newman and Girvan (2004) and Newman (2006). Table 2 provides a list of these topics. Of each article j is determined which fraction of the words fall into topic i, which is denoted as $f_{i,j}$ (note that $\sum_i f_{i,j} = 1$). Monthly country-level topic frequencies *f[Topic]* are computed by taking the average of the frequencies of all articles of that month.

The sentiment (*Sent*) of article j is assessed using the following formula:

$$Sent_j = \frac{POS_j - NEG_j}{a_j} \tag{3.1}$$

Where POS_j, NEG_j and a_j denote the number of positive, negative and total words in the article, respectively. Note the difference here with the sentiment variable defined by Tetlock et al. (2008), which only focused on negative words. To examine whether, for instance, a sentiment in Government related articles has a different impact than in articles focusing on Macroeconomics, an article's sentiment is decomposed into a topic-specific sentiment measure as follows:

$$Sent_{i,j} = f_{i,j} * Sent_j \tag{3.2}$$

Where $Sent_{i,j}$ is the sentiment of article j specific to topic i. To illustrate: an article with a sentiment measure of -10% that contains for 80% markets-related words gets a markets-specific sentiment score of -8% (note again that $\sum_i Sent_{i,j} = 1$).

Lastly, the authors investigate whether article sentiment has a stronger explanatory value when combined with the *entropy* (unusualness) of an article. The intuition behind this is that positive/negative sentiment may have a bigger impact on returns if it occurs in months with unusual news (as measured by the language). The variable $SentEnt_{i,j}$ captures the sentiment specific to topic i of article j interacted with its entropy score. Just as with topic frequencies, country-level monthly figures are obtained for *Sent* and *SentEnt* for each topic by taking the average of the article-level values of that month.

Table 3 shows the descriptive statistics from the data that I have obtained. This paper uses the exact same data sources as Calomiris and Mamaysky (2019) and stuck to their description of how they obtained the data as closely as possible. Occasionally, multiple data series fitted their description, in which case I used my own judgment to select the best fit. Comparing the descriptive statistics of this paper to the numbers reported by the authors shows that they are very similar. This is an indication that I have extracted the right data from the right databases and as a result one would expect my panel regressions to be similar to those of the authors as well.

Table 3: Descriptive Statistics

Descriptive statistics for the full sample period April 1998 to December 2015. For each variable the mean, standard deviation, 5^{th} percentile, 95^{th} percentile and the one-lag autocorrelation coefficient *AR(1)* are displayed. The *AR(1)* coefficient is obtained for each country individually and then averaged. *N* is the total number of observations for all countries together. Descriptions of all variables and their sources can be found in Table 2.

Emerging markets						
	mean	sd	5%	95%	AR(1)	N
return	1.014	9.359	-13.540	15.329	0.146	5549
$return^{12}$	17.610	43.109	-42.110	91.835	0.926	5241
sigma	25.633	16.896	9.630	55.580	0.552	5552
value	0.906	0.879	0.195	2.321	0.975	4983
gdp	3.901	4.680	-4.239	10.232	0.937	4881
gdpdeflator	6.818	10.212	-1.160	23.374	0.929	4536
cp	60.133	45.223	12.416	146.480	0.980	5544
rate	8.198	9.415	0.924	21.236	0.981	4742
dexch	0.406	4.022	-4.389	5.741	0.135	5938
entropy	2.429	0.171	2.111	2.649	0.787	5964
artcount	26.022	30.146	3.200	73.387	0.769	5964
sMkt	-0.005	0.002	-0.009	-0.002	0.687	5964
fMkt	0.334	0.080	0.195	0.460	0.643	5964
sGovt	-0.007	0.004	-0.016	-0.002	0.566	5964
fGovt	0.290	0.096	0.163	0.480	0.597	5964
sCorp	-0.002	0.001	-0.004	-0.001	0.623	5964
fCorp	0.169	0.034	0.124	0.229	0.578	5964
sComms	-0.002	0.001	-0.004	-0.001	0.486	5964
fComms	0.159	0.064	0.082	0.289	0.691	5964
sMacro	-0.001	0.001	-0.002	-0.000	0.622	5964
fMacro	0.048	0.028	0.025	0.105	0.611	5964

Developed markets						
	mean	sd	5%	95%	AR(1)	N
return	0.574	6.654	-10.985	10.605	0.121	4714
$return^{12}$	8.830	27.624	-40.017	51.512	0.931	4462
sigma	21.839	12.373	9.710	45.920	0.661	4725
value	0.846	0.532	0.279	1.693	0.983	4498
gdp	2.087	3.279	-3.550	6.541	0.944	4473
gdpdeflator	1.812	2.268	-1.332	5.379	0.930	4473
cp	117.653	38.529	60.594	187.047	0.984	4395
rate	4.003	2.173	1.073	6.250	0.987	4897
dexch	0.020	2.714	-4.423	4.356	0.067	4686
entropy	2.455	0.170	2.124	2.666	0.842	4899
artcount	106.700	214.462	12.733	316.968	0.548	4899
sMkt	-0.004	0.002	-0.007	-0.002	0.746	4899
fMkt	0.350	0.055	0.270	0.450	0.713	4899
sGovt	-0.004	0.002	-0.007	-0.001	0.617	4899
fGovt	0.237	0.055	0.150	0.328	0.610	4899
sCorp	-0.002	0.001	-0.003	-0.001	0.636	4899
fCorp	0.220	0.050	0.159	0.314	0.751	4899
sComms	-0.000	0.000	-0.001	-0.000	0.464	4899
fComms	0.027	0.013	0.012	0.052	0.555	4899
sCredit	-0.002	0.001	-0.004	-0.001	0.681	4899
fCredit	0.167	0.023	0.137	0.208	0.576	4899

There are a few aspects to note when looking at the data. First of all, one can see that Emerging Markets are riskier and achieve higher average returns as a result, which is in line with finance theory. Furthermore, emerging markets are characterized by higher GDP growth, inflation, and interest rates. Looking at the news variables, it becomes clear that there are roughly four times as many articles published on developed markets compared to emerging markets, which is understandable given the relatively larger size of the stock markets in developed economies. The average sentiment for each of the topics is very similar in both markets and topic frequencies as well, with the frequency of commodities-related articles being the exception, which is a lot higher in emerging market countries. One explanation for this exception could be that the stock prices of emerging markets have historically correlated strongly with prices of commodities, making news in this field highly relevant (Basher, Haug, & Sadorsky, 2012).

Finally, the very high one-lag autocorrelation coefficients *AR(1)* for $return^{12}$ are a direct consequence of the overlapping return problem, as two subsequent data points have 11 of the 12 monthly returns used in the computation in common. The high autocorrelation in *value* has a similar explanation, since it is obtained using the average index level of 4.5 years to 5.5 years ago, which again includes 11 similar observations in two consecutive months. High autocorrelation for other variables is partially due to the fact that for some of them quartertly (*gdp*, *gdpdeflator*) or yearly (*cp*) data is used, yielding identical values for 3 or 12 consecutive months.

4 Methodology

4.1 Replication of Calomiris and Mamaysky (2019)

The core regression employed by Calomiris and Mamaysky (2019) is a panel regression with country fixed effects. The authors do not specify the regression equation, but it looks as follows:

$$r_{i,t+1} = \alpha + X_{i,t} \cdot \beta + \mu_i + v_{i,t+1}, \tag{4.1}$$

where $r_{i,t+1}$ is the (month or 12-month) stock return for country i in month *t+1* and $X_{i,t}$ are the lagged values of the regressors. The fixed effect μ_i captures all time-invariant country specific characteristics, such that the regression is not impacted or biased by unobserved individual influences.

Since the authors did not find significant forecasting power for one-month ahead returns, they focus on panel regressions with 12-month returns as the independent variable, i.e. $r_{i,t+1} = return^{12}_{i,t+1}$. However, this paper also performs regressions of monthly returns, in which case $r_{i,t+1}$ is equal to $return_{i,t+1}$.

In order to assess the additional explanatory value of the news variables over the macro and return regressors, a baseline regression will first be employed where $X'_{i,t} = (sigma_{i,t}\ sigma_{i,t-1}\ return_{i,t}\ return_{i,t-1}\ value_{i,t}\ gdp_{i,t}\ gdpdeflator_{i,t}\ cp_{i,t}\ rate_{i,t}\ dexch_{i,t})$.

For the regression including news variables, 1-month lags of *artcount*, *entropy* and the topic-specific frequencies and sentiments are added to the regressors. Furthermore, panel regressions are performed in which the sentiment variables are replaced by the sentiment-entropy variables.

The authors have found that the global financial crisis has had a big impact on sentiment for each topic. For example, the sentiment score of market-related articles was more positive than government-related articles prior to 2007 and the opposite holds true for the years thereafter. In order to take possible structural breaks in the regression coefficients into account, the panel regressions are performed for two subperiods: April 1998 - February 2007 and March 2007 - December 2015. All in all, this yields 9 panel regressions for developed and emerging markets, i.e. *Base*, *Sent* and *SentEnt* regressions for the entire period as well as the two subperiods.

Autocorrelation is an often encountered issue when dealing with panel data that can have a big impact as it leads to significant underestimation of the standard errors when ignored. Calomiris and Mamaysky (2019) cluster their standard errors in a response to this issue, but they do not motivate this choice, nor do they explain why they sometimes cluster the standard errors by time and sometimes by time and country. Therefore, I will do research on the data in order to determine which method is most suitable to deal with possible auto correlation. Figure 1 shows the average residual for each month in the developed market panel regression for 12-month returns. As one can observe, the residuals seem to correlate in subsequent months. Taking the average residual by country shows large variations as well, ranging from -55 to 39. This suggests that both time effects and country effects are present in the data.

In order to confirm whether the suspected time and country effects exist, I use a method proposed by Petersen (2009). When standard errors clustered by country and time are much larger (at least three times) than the standard errors clustered by only one dimension, both effects are present. When this is the case, Petersen (2009) shows that double clustering (i.e. by time and country) yields unbiased standard errors, whereas often used Fama-Macbeth (Fama and MacBeth, 1973) and Newey-West (Newey and West, 1987) procedures do not. Applying Petersen (2009), I find that country and time effects are present in both emerging and developed markets for the entire period as well as the subperiods. Hence, I will use double-clustered standard errors for all panel regressions. The double-clustered errors are White standard errors adjusted to account for the possible correlation within clusters and are also called Rogers standard errors in financial literature (Rogers, 1993). I use software provided by Correia (2016) to implement

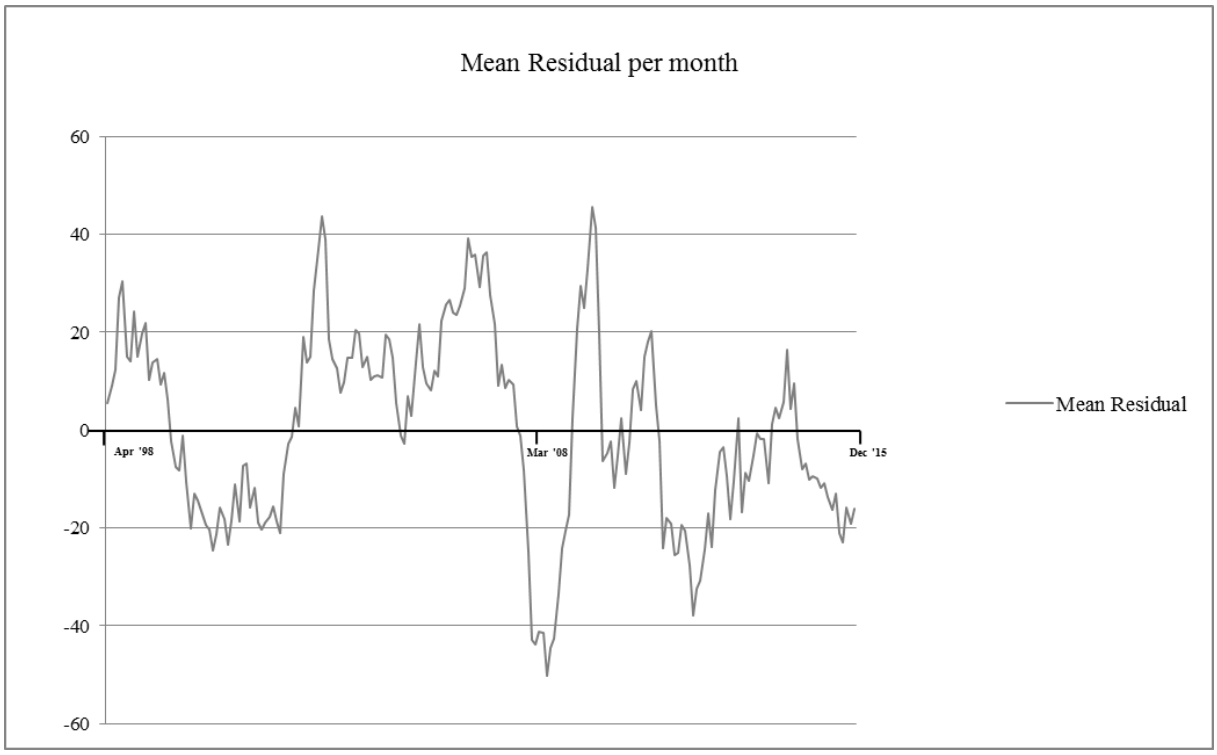

Figure 1: Mean residual for each month in the panel regression with fixed effects for developed markets 12-month returns.

the double-clustering in *Stata*.

4.2 Ordinary Least Squares regressions by country

The first extension that this paper will add to the existing literature is investigating the forecasting power of news variables for monthly and 12-month returns on country level using Ordinary Least Squares (OLS) regressions. The regression equation looks similar to (4.1), except for the obvious fact that the country-specific term μ is not necessary here.

$$r_{t+1} = \alpha + X_t \cdot \beta + \epsilon_{t+1}, \tag{4.2}$$

where r_{t+1} and X_t are equal to the specifications in (4.1). Since I am no longer dealing with panel data, the Newey-West procedure provides robust standard errors in the presence of autocorrelation and heteroskedasticity (Newey and West, 1987).[2] The baseline regression will also be performed for each country, as the increase in *R-squared* when the news variables are included is of interest.

[2]In line with Greene (2003), the lag length is set to the integer part of $N^{1/4}$, where N is the number of observations.

4.3 Least Angle Regressions by country

Another way to evaluate the importance of the news variables in the regressions is by employing a "lasso" regression (Tibshirani, 1996). This method minimizes the sum of squared residuals with the constraint that the sum of the absolute value of the (standardized) coefficients is less than a constant s, i.e.

$$\begin{aligned} \text{minimize} \quad & \sum_{t=1}^{T} \left(r_{t+1} - \sum_{j} X_{t,j} \cdot \beta_j \right) \\ \text{subject to} \quad & \sum_{j} \beta_j \leq s \end{aligned} \tag{4.3}$$

If the value of s is sufficiently strict, part of the coefficients will be set to zero in the lasso regressions. Hence, it is of interest to see how many of the news variables retain non-zero coefficients when the total absolute value of all coefficients is limited. Tibshirani (1996) proposes several ways to do determine the value of s, for instance by minimizing the prediction error. I have determined s by estimating the prediction error using cross-validation as described in Efron and Tibshirani (1994) and subsequently picking the value of s that minimizes the prediction error. However, in the subsequent lasso regressions this value of s was not strict enough the force any of the coefficients, news variables and other variables, to zero. As the purpose in this application is to see whether news variables or other variables survive when introducing lasso restrictions, this method is not desirable. Alternatively, one can set s to a certain percentage of the sum of the absolute value of the coefficients in the unrestricted OLS regression. In this paper, lasso regressions will be performed for s equal to 30%, 50%, and 70% of the sum of absolute coefficients in the unrestricted OLS regressions. The importance of news variables relative to other variables can then be assessed by determining what percentage of the variables in each group is non-zero.

In order to save computation time, a more efficient alternative for the lasso algorithm will be employed, called the Least Angle Regression (LARS) algorithm (Efron et al., 2004). The algorithm works as follows:

1. Initially, all coefficients are equal to zero.
2. Select the variable x_j most correlated with the returns r.
3. Increase the coefficient of this variable in the direction of the correlation. The residuals $e = (r_{t+1} - \sum_j X_{t,j} \cdot \beta_j)$ are updated along the way. The increasing of the variable stops when another variable x_k has as much correlation with the residuals e as x_j.

4. Increase (b_j, b_k) in the direction of their joint least squares, until another variable x_m has as much correlation with the (updated) residual e.

5. Continue until all predictors are in the model.

Efron et al. (2004) show that if one removes a variable with a non-zero coefficient from the set of variables when it hits zero, the LARS algorithm produces the entire path of lasso solutions as s increases from 0 to infinity. This paper will use MATLAB code provided by Efron et al. (2004) to apply the algorithm.

4.4 Weighted Least Squares regressions by country

Although the OLS with the Newey-West procedure yields non-biased standard errors, it is not necessarily the most efficient. For example, in periods where realized returns are very volatile (e.g. crises), these returns form noisy proxies for expected returns which makes equal weighting in OLS inefficiently high. In response, Johnson (2018) assesses return predictability using a Weighted Least Squares (WLS) estimator that is around 25% more efficient than OLS because it incorporates time-varying volatility into its point estimates.

Overlapping data is an important issue that Calamiris and Mamaysky (2019) seem to have overlooked. One-year returns ($returns^{12}$) are obtained by accumulating the next twelve monthly returns, causing two subsequent one-year returns to have eleven monthly returns in common. The overlapping of observations creates a moving average (MA) error term which renders OLS parameter estimates inefficient and hypothesis tests biased (Hansen and Hodrick, 1980). Using Monte Carlo simulations, Harri and Brorsen (1998) confirmed the occurrence efficiency loss and biased test statistics for a several distributions of the dependent variable, even when Newey-West standard errors were used. Johnson (2018) not only provides more efficient estimation techniques than OLS, but also proposes a procedure to efficiently handle the overlapping return problem.

The estimator of Johnson (2018), weighted least squares using ex ante variance (WLS-EV), scales regression residuals by an estimate of ex ante return volatility. In this paper, the techniques of Johnson (2018) will be applied to assess the predictability of one-year returns using news variables when overlapping returns and heteroskedasticity have been accounted for efficiently. The following regression is estimated:

$$r_{t+1,t+12} = X_t \cdot \beta + \epsilon_{t+1,t+12}, \tag{4.4}$$

where $r_{t+1,t+12}$ denotes the one-year ahead return and X_t contains a constant and the

monthly regressors. However, I cannot apply WLS-EV directly to the above equation, since overlapping returns need to be dealt with first. Johnson (2018) makes use of insights in Hodrick (1992), who showed that overlapping regressions of 12-month returns $r_{t+1,t+12}$ can be mapped to the following equivalent non-overlapping regressions of monthly returns:

$$r_{t+1} = \bar{X}_t \cdot \beta + \epsilon_{t+1} \tag{4.5}$$

where $\bar{X}_t$ is equal to the rolling sum of past X_t, i.e. $\sum_{s=0}^{s=11} X_{t-s}$. Johnson (2018) proves that after scaling, the obtained coefficients from the regressions in (4.4) and (4.5) both converge to the true beta, but that the latter provides far more accurate standard errors than the overlapping regression.

The estimate of β is obtained in three steps.

1. Estimate σ_t^2, the conditional variance of next-month unexpected returns ϵ_{t+1}.

2. Estimate $\hat{\beta}_{WLS-EV}$ using:

$$\hat{\beta}_{WLS-EV} = \arg\min_{\beta} \sum_{t=1}^{T} \left(\frac{r_{t+1} - \bar{X}_t \cdot \beta}{\hat{\sigma}_t} \right) \tag{4.6}$$

   where $\hat{\sigma}_t$ is obtained by taking the square root of the estimate of σ_t^2 from Step 1. This estimator can be implemented by performing an OLS regression of $\frac{r_{t+1}}{\hat{\sigma}_t}$ on $\frac{\bar{X}_t}{\hat{\sigma}_t}$.

3. Scale the resulting $\hat{\beta}_{WLS-EV}$ and standard errors by $E_T(X_t' X_t)^{-1} E_T(\hat{X}_t{}' \hat{X}_t)$. Here E_T denotes the sample average. In a univariate scenario, this would be equivalent to scaling by the ratio of the variance of $\bar{X}_t$ and the variance of X_t. This step needs to be performed in order to assure that the obtained coefficients from the regression of $\bar{X}_t$ correspond to the coefficients of the regression that I originally aimed to estimate.

There are various ways to estimate σ_t^2, the conditional variance of next-month returns. I will use prior-month and prior-year realized variance (RV) to predict next-month variance, as these are the strongest predictors according to Johnson (2018). Hence, the estimate for σ_t^2 will be obtained using the following equation:

$$\hat{\sigma}_{t+1}^2 = \hat{a} + \hat{b} \cdot RV_t + \hat{c} \cdot RV_{t-11,t}, \tag{4.7}$$

where $\hat{a}$, $\hat{b}$, and $\hat{c}$ are estimated in a regression of RV_{t+1} on a constant, RV_t, and $RV_{t-11,t}$. The monthly RV is computed as the sum of squared daily log market returns in that month, and the yearly RV is obtained by taking the average of twelve underlying monthly RVs.

The third research question is answered by using U.S. data for the explanatory variables for the WLS regressions of all countries.

5 Results

Table 4 reports a summary of the results for all regressions performed. In the subsections that follow, I comment on each of the regressions individually.

Table 4: Summary of the results for all regressions
Summary of the results for all regressions described in section 4. Results are reported for 12-month returns and monthly returns, respectively. *Panel* refers to the panel regressions replicated from Calomiris and Mamaysky (2019). *OLS* and *LARS* display the country-level Ordinary Least Squares regressions and Least Angle Regressions. *WLS* refers to the country-level Weighted Least Squares regressions accounting for overlapping returns as proposed by Johnson (2018). *WLS - US* reports these regressions when U.S. variables are used to explain country-level returns. *Sign. News Variables* reports the number of news variables that are significant at the 5% level. LARS in an exception, here the number of non-zero news variables is reported. A list of these news variables can be found in Table 2. *R2* reports the *R-squared* of the regression including news variables, while *R2 Base* reports the equivalent when news variables are excluded. *R2 increment* is the increase in *R-squared* when news variables are included, which is also given in % in *increment in %*. For the country-level regressions, the numbers are averages.

12-month returns

Regression	Countries	Sign. News Variables	R2	R2 Base	R2 Increment	increment in %
Panel	DM	2	0.324	0.262	0.062	24%
	EM	5	0.218	0.129	0.089	69%
OLS	DM	2.6	0.633	0.472	0.161	34%
	EM	3.0	0.569	0.420	0.149	38%
LARS	DM	8.1	0.613	0.472	0.141	30%
	EM	8.4	0.559	0.420	0.139	33%
WLS	DM	2.1	0.241	0.099	0.142	143%
	EM	2.8	0.240	0.157	0.083	53%
WLS - US	DM	1.9	0.270	0.111	0.159	143%
	EM	2.4	0.262	0.177	0.134	76%

Monthly returns

Regression	Countries	Sign. News Variables	R2	R2 Base	R2 increment	Increment in %
Panel	DM	2	0.055	0.042	0.013	31%
	EM	3	0.039	0.029	0.010	34%
OLS	DM	1.1	0.169	0.085	0.084	99%
	EM	1.0	0.213	0.127	0.086	68%
LARS	DM	6.0	0.143	0.085	0.058	68%
	EM	7.7	0.178	0.127	0.051	40%

5.1 Results of the Replication of Calomiris and Mamaysky (2019)

Table 5 and 6 report the 12-month return and monthly return panel regressions for developed markets, respectively. The equivalent regressions for emerging markets are reported in the appendix. *Stata* code used for the replication is included in the appendix as well.

Table 5: Developed markets forecasting panel for 12-month returns

Panel regressions for developed market 12-month returns ($return^{12}$). Significant coefficients at the 1%, 5%, and 10% level are labeled with "***", "**", and "*", respectively. Descriptions of the variables in the regressions can be found in Table 2. The *Base*, *Sent*, and *SentEnt* columns report the baseline regression, sentiment regression, and sentiment-entropy regression for the entire period and two subperiods. The *stderr* row indicates how the standard errors are clustered, where *both* means double-clustered by country and time. The absolute values of the coefficients of the news variables are larger in the replication than in the original paper. This is due to the fact that this paper standardizes the news variables by subtracting the mean and dividing by the standard deviation, whereas the authors did not subtract the mean.

	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	0.158	0.243	0.252	0.028	0.081	0.086	0.331*	0.401**	0.392**
$sigma_{t-2}$	0.057	0.115	0.112	0.113	1.137	0.140	0.224	0.288	0.279
$return_{t-1}$	0.284	0.127	0.129	-0.251	-0.264	-0.266	0.277	0.115	0.121
$return_{t-2}$	0.031	-0.097	-0.097	-0.190	-0.200	-0.199	0.009	-0.059	-0.056
$value_{t-1}$	29.000***	32.315***	32.482***	22.539***	23.198***	23.296***	28.984***	30.682***	30.878***
gdp_{t-1}	-0.611	-1.134*	-1.175*	0.946	0.882	0.854	-1.886***	-1.821***	-1.809***
$gdpdeflator_{t-1}$	1.112	0.613	0.559	0.913*	0.922**	0.922*	-1.210	-0.947	-0.970
cp_{t-1}	-0.364***	-0.302***	-0.304***	-0.024	-0.033	-0.033	-0.471***	-0.465***	-0.466***
$rate_{t-1}$	-4.699***	-5.187***	-5.206***	-19.782***	-18.527***	-18.399***	-4.338***	-4.020***	-4.081***
$dexch_{t-1}$	0.751	0.811*	0.820*	-1.033**	-0.882*	-0.884*	1.340**	1.230**	1.226**
$entropy_{t-1}$		-2.062	1.221		-33.029	-31.065		-4.219	-4.428
$artcount_{t-1}$		-1.697	-2.989		-7.200	-8.157		-6.303	-3.804
$sMkt_{t-1}$		7.116**	7.603**		-0.142	-0.364		7.977**	7.849**
$fMkt_{t-1}$		6.444*	6.537*		-0.594	-0.538		4.264	4.522
$sGovt_{t-1}$		-5.137**	-5.336**		-1.248	-1.529		-4.903*	-4.883*
$fGovt_{t-1}$		1.507	1.568		-1.409	-1.486		-1.460	-1.059
$sCorp_{t-1}$		-3.581*	-2.997		1.731	2.679		-5.010**	-4.780**
$fCorp_{t-1}$		-3.579	-3.186		-4.835	-4.312		1.455	1.943
$sComms_{t-1}$		0.825	0.958		2.149	1.873		1.606	1.611
$fComms_{t-1}$		2.583	2.713		3.435	3.266		2.445	2.729
$sCredit_{t-1}$		4.388*	3.553		-0.295	-0.860		1.136	0.501
$fCredit_{t-1}$		omitted	omitted		omitted	omitted		omitted	omitted
R2	0.262	0.3236	0.3233	0.4754	0.5016	0.5028	0.4878	0.4998	0.5089
start	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4118	4118	4118	1876	1876	1876	2242	2242	2242
stderr	both	both	both	both	both	both	both	both	both

Table 6: Developed markets forecasting panel for monthly returns
Panel regressions for developed market monthly returns (*return*). Significant coefficients at the 1%, 5%, and 10% level are labeled with "***", "**", and "*", respectively. Descriptions of the variables in the regressions can be found in Table 2. The *Base*, *Sent*, and *SentEnt* columns report the baseline regression, sentiment regression, and sentiment-entropy regression for the entire period and two subperiods. The *stderr* row indicates how the standard errors are clustered, where *both* means double-clustered by country and time. The absolute values of the coefficients of the news variables are larger in the replication than in the original paper. This is due to the fact that this paper standardizes the news variables by subtracting the mean and dividing by the standard deviation, whereas the authors did not subtract the mean.

	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	-0.017	0.004	0.006	0.112**	0.119**	0.118**	-0.062	-0.041	-0.039
$sigma_{t-2}$	0.001	0.016	0.016	-0.124*	-0.119	-0.120	0.074	0.093	0.093
$return_{t-1}$	0.147*	0.113	0.115	0.151**	0.146**	0.148**	0.108	0.053	0.054
$return_{t-2}$	-0.032	-0.050	-0.050	-0.094	-0.098	-0.098	-0.001	-0.033	-0.034
$value_{t-1}$	2.190**	2.657***	2.699***	2.061**	2.021*	1.991*	2.108	2.644**	2.772**
gdp_{t-1}	0.018	-0.038	-0.046	0.087	0.055	0.057	-0.004	-0.045	-0.052
$gdpdeflator_{t-1}$	-0.012	-0.080	-0.085	0.124	0.137	0.137	-0.230*	-0.281*	-0.283*
cp_{t-1}	-0.016	-0.014	-0.014	-0.008	-0.010	-0.009	-0.029	-0.042*	-0.043*
$rate_{t-1}$	-0.308**	-0.415***	-0.427	-1.528***	-1.370**	-1.363**	-0.253	-0.356*	-0.379*
$dexch_{t-1}$	0.100	0.094	0.097	0.252	0.251	0.252	0.024	0.017	0.019
$entropy_{t-1}$		2.641	3.246		-3.217	-3.200		5.973*	6.590*
$artcount_{t-1}$		0.586	0.480		-0.015	-0.045		0.598	0.576
$sMkt_{t-1}$		1.564***	1.541***		0.309	0.089		2.411	2.500***
$fMkt_{t-1}$		0.912	0.784		-1.137	-1.236		1.736	1.599
$sGovt_{t-1}$		-1.339**	-1.387**		-0.068	-0.057		-1.938**	-2.001**
$fGovt_{t-1}$		-0.689	-0.750		-0.529	-0.533		-1.000	-1.057
$sCorp_{t-1}$		0.221	0.378		0.284	0.426		0.499	0.610
$fCorp_{t-1}$		-0.187	-0.197		-0.956	-0.913		0.530	0.466
$sComms_{t-1}$		0.116	0.085		0.093	0.052		0.152	0.099
$fComms_{t-1}$		0.435	0.040		-0.419	-0.432		0.958	0.855
$sCredit_{t-1}$		-0.028	-0.111		-0.256	-0.249		-0.557	-0.664
$fCredit_{t-1}$		omitted	omitted		omitted	omitted		omitted	omitted
R2	0.042	0.055	0.066	0.106	0.114	0.114	0.059	0.093	0.095
start	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4118	4118	4118	1876	1876	1876	2242	2242	2242
stderr	both	both	both	both	both	both	both	both	both

Looking at the baseline regressions of the developed markets, the results are very similar. In my research as well as in that of Calamiris and Mamaysky (2019), *value*, *cp* and *rate* are significant at the 1% level in all (sub)samples. Slight differences exist in the baseline regressions of the subperiods, where occasionally one of the variables is significant at the 10% level in my regression and not significant in the regression of the authors. This is most likely due to slight differences in the data. For emerging markets, the baseline regressions are very similar as well, but here the volatility lags are the most significant predictors of returns.

One surprising finding is the inverse relationship between GDP growth and 12-month returns in the post global crises period (significant at the 1% level for both country groups in the replication). Although this may seem counter-intuitive at first, there is a plausible economic explanation. It has been established that expansionary monetary policy has a positive impact on stock returns (Thorbecke, 1997). When deciding on the monetary policy, the growth rate of GDP is an important factor for central banks such as the Federal Reserve and the European Central Bank. Hence, high GDP growth rates may foreshadow contractionary monetary policy, which negatively impacts stock returns.

Looking at the monthly return regressions, there is indeed low forecasting power, as Calomiris and Mamaysky (2019) claimed, with *R-squared* ranging between 0.020 and 0.059.

Next, I turn to the replicated regressions containing the news variables. The authors conclude that news variables have explanatory value mainly by pointing at the increase in *R-squared* when the variables are added to the baseline model. The absolute increase in *R-squared* for the full-sample regressions is slightly higher for the replication than for the authors' results (DM: 0.062 vs. 0.051, EM: 0.089 vs. 0.057). However, since the initial baseline values were also higher for the replication, the proportional increase is similar. Hence, the replication confirms the authors claim that including word flow measures increases the *R-squared*. On a critical note, the inclusion of 12 news variables adds considerably to the complexity of the model and one may argue that the increase in *R-squared* is not sufficiently large to justify the use of news variables (for some subperiods the increment is less than 0.03). Including entropy in the sentiment measures has no significant impact on the *R-squared* in the replication, which is in line with the authors' findings.

Although the authors have shown that the *R-squared* increases when including the news variables, individual news variables were rarely significant in the developed country group. For the entire period regression, only *fCorp* and *sCredit* were significant at the 10% level in Calomiris and Mamaysky (2019). The replication reports more significant variables in the full period, with significant coefficient for markets and government sentiment at the 5% level. Surprisingly, the coefficient for government sentiment is negative, which means that a negative sentiment score

has a positive impact on returns. One possible explanation is that a negative sentiment does not necessarily mean bad news. Sometimes, negative sentiment can correspond to good news if the sentiment is describing problems that a government (or corporation) is trying to address. In the early subperiod, no variables are significant in both researches.[3] In the second subperiod several topic-specific sentiment variables are significant at 5% and 10% levels in both researches. Lastly, the replication is not able to confirm the authors' finding of a significant positive coefficient for *entropy* in the second subperiod. The largest difference between the regressions of developed and emerging markets is that the topic-specific frequencies rather than the topic-specific sentiments are often significant. Furthermore, the significant coefficients are all negative. It could be the case that the number of articles on emerging economies mainly increases in times of bad news, which explains the negative relationship between frequencies and returns. Furthermore, it can explain why the sentiment variables are not significant, as their impact is already captured by the frequency variables.

The most unexpected finding that I came across when performing the replication was that *Stata* omitted the final variable *fCredit* (for DM) and *fMacro* (for EM) in regressions because of collinearity. As the topic-specific frequencies always add up to 1, the existence of collinearity makes sense, as one can calculate what the frequency of the final topic is by looking at the frequencies of the other topics. In order to identify this dependence, I have regressed *fCredit* on the other regressors in the *Sent* regression of developed countries 12-month returns.[4] As can be found in Table 7, the *R-squared* of this regression is 0.952, meaning that the other regressors are able to explain over 95% of the variation in *fCredit*. Furthermore, all frequency variables have negative coefficients with p-values of 0.000, confirming my hypothesis. Hence, it is understandable that *Stata* omits the variable because of collinearity. The authors have probably used different software that does not automatically omit variables, but this does not mean that the issue is not present and it has most likely impacted the significance of individual variables.

To summarize, the replication confirms the authors' claim that adding news variables increases the *R-squared*. For the full sample and 12-month returns, *R-squared* increments of 24% and 69% are found for DM and EM, respectively. For monthly returns the *R-squared* increments are 31% (DM) and 34% (EM), but in absolute terms very small (0.013 for DM and 0.010 for EM). The replication finds that between 2 (DM, 12-month returns and monthly returns) and 5 (EM, 12-month returns) news variables are significant at the 5% level. This is comparable to Calomiris and Mamaysky (2019), but one could still argue that it is rather poor given that 12

[3]There is one exception: *sCorp* is significant at the 10% level in the *SentEnt* regression of the authors.
[4]Similar results are obtained for *fMacro* in the emerging countries group.

news variables were added to the model.

Table 7: Demonstration of collinearity
Regression of *fCredit* on the other variables in the developed markets panel regression. The corresponding panel regressions in which *fCredit* was omitted are reported in Table 5 (12-month returns) and Table 6 (monthly returns).

	Coefficient	Standard error	p-value
return	0.000	0.001	0.963
sigma	0.000	0.000	0.512
value	0.021	0.008	0.009
gdp	0.007	0.001	0.000
gdpdeflator	-0.003	0.002	0.072
cp	0.000	0.000	0.926
rate	0.007	0.002	0.001
dexch	0.003	0.001	0.042
entropy	-0.219	0.031	0.000
artcount	-0.073	0.025	0.004
sMkt	-0.003	0.008	0.738
fMkt	-2.203	0.012	0.000
sGovt	0.007	0.007	0.355
fGovt	-2.196	0.012	0.000
sCorp	-0.039	0.006	0.000
fCorp	-2.027	0.011	0.000
sComms	0.189	0.004	0.000
fComms	omitted		
sCredit	-0.076	0.008	0.000
R-squared	0.952		

5.2 Results of OLS regressions by country

The results of the OLS regressions by country are summarized in Table 8 and Table 9 for 12-month returns and monthly returns of the DM group, respectively. The appendix reports similar results for the EM group.

In the 12-month return regressions, on average 2.6 of the 10 news variables are significant at the 5% level in developed market countries, and 3.0 of the 10 in emerging countries. These statistics are roughly in line with the results from the panel regressions. However, there is a large variation in the number of variables being significant per country. In some countries none of the news variables is found to be significant (Finland, Kenya, and Nigeria), whereas in others 6 or 7 variables yield significant coefficients (Australia, Great-Britain, and Slovakia).

Table 8: Developed markets country-level OLS for 12-month returns

Country-level Ordinary Least Squares (OLS) regressions for developed market countries of next 12-month returns ($return^{12}$). The same variables have been used as in the panel regression reported in Table 5, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "+" ("-") denotes a positive (negative) coefficient significant at the 5% level. *Total* reports the total number of significant news variables. "R2" and "R2 increment" denote the *R-squared* of the regression including news variables and the increment over the corresponding baseline (regression without news variables), respectively. The standard errors are obtained using the Newey-West procedure (Newey and West, 1987).

	AU	AT	BE	CA	DK	FI	FR	DE	GR	IE	IT	JP
$entropy_{t-1}$		+										-
$artcount_{t-1}$		-			-							
$sMkt_{t-1}$	+	+	+	+			+			+		
$fMkt_{t-1}$												
$sGovt_{t-1}$	-	-		-				-	-			
$fGovt_{t-1}$	-			-			-	-				
$sCorp_{t-1}$			-									
$fCorp_{t-1}$	-	-		-							+	
$sComms_{t-1}$												
$fComms_{t-1}$	-				+							
$sCredit_{t-1}$			+	+					+		+	
$fCredit_{t-1}$	-								+			
Total	6	5	3	5	2	0	2	2	3	1	2	1
R2	0.650	0.598	0.718	0.677	0.465	0.511	0.697	0.732	0.617	0.712	0.777	0.411
R2 increment	0.455	0.321	0.129	0.166	0.122	0.039	0.061	0.116	0.057	0.069	0.171	0.135
Nobs	211	179	161	139	175	179	179	179	176	179	134	211

	LU	NL	NZ	NO	PT	SG	ES	SE	CH	GB	US
$entropy_{t-1}$			-					-			
$artcount_{t-1}$									-	-	
$sMkt_{t-1}$	+	+		+			+			+	
$fMkt_{t-1}$						+			+		
$sGovt_{t-1}$					-					-	-
$fGovt_{t-1}$										-	
$sCorp_{t-1}$					-		-	+			
$fCorp_{t-1}$	+		-	-						-	+
$sComms_{t-1}$										+	
$fComms_{t-1}$	-										
$sCredit_{t-1}$									+		
$fCredit_{t-1}$										+	
Total	3	1	2	2	2	1	2	2	3	7	2
R2	0.634	0.701	0.605	0.609	0.688	0.643	0.707	0.536	0.696	0.650	0.544
R2 increment	0.259	0.077	0.262	0.215	0.115	0.141	0.145	0.096	0.184	0.177	0.187
Nobs	138	179	187	174	179	132	179	211	211	211	211

Table 9: Developed markets country-level OLS for monthly returns
Country-level Ordinary Least Squares (OLS) regressions for developed market countries of monthly returns. The same variables have been used as in the panel regression reported in Table 6, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "+" ("-") denotes a positive (negative) coefficient significant at the 5% level. *Total* reports the total number of significant news variables. "R2" and "R2 increment" denote the *R-squared* of the regression including news variables and the increment over the corresponding baseline (regression without news variables), respectively. The standard errors are obtained using the Newey-West procedure (Newey and West, 1987).

	AU	AT	BE	CA	DK	FI	FR	DE	GR	IE	IT	JP
$entropy_{t-1}$	+				+							
$artcount_{t-1}$							+	+				
$sMkt_{t-1}$	+	+						+			+	+
$fMkt_{t-1}$										+		
$sGovt_{t-1}$								-				
$fGovt_{t-1}$	-							-				
$sCorp_{t-1}$						+						
$fCorp_{t-1}$	-											
$sComms_{t-1}$												
$fComms_{t-1}$												
$sCredit_{t-1}$												
$fCredit_{t-1}$												
Total	4	1	0	0	1	1	1	4	0	1	1	1
R2	0.141	0.164	0.187	0.168	0.203	0.206	0.156	0.225	0.141	0.234	0.288	0.165
R2 increment	0.110	0.075	0.091	0.047	0.092	0.054	0.078	0.137	0.050	0.09	0.134	0.112
Nobs	211	179	161	139	175	179	179	179	176	179	134	211

	LU	NL	NZ	NO	PT	SG	ES	SE	CH	GB	US
$entropy_{t-1}$											
$artcount_{t-1}$						+				+	+
$sMkt_{t-1}$			+								
$fMkt_{t-1}$											
$sGovt_{t-1}$			-					-	-		
$fGovt_{t-1}$							-				
$sCorp_{t-1}$											
$fCorp_{t-1}$	+									-	
$sComms_{t-1}$											
$fComms_{t-1}$											
$sCredit_{t-1}$											
$fCredit_{t-1}$											
Total	1	1	1	0	0	1	2	1	0	2	1
R2	0.208	0.127	0.141	0.178	0.152	0.248	0.139	0.159	0.133	0.181	0.134
R2 increment	0.081	0.057	0.07	0.1	0.036	0.111	0.080	0.089	0.050	0.131	0.077
Nobs	138	179	187	174	179	132	179	211	211	211	211

The *R-squared* are a lot higher in the country regressions than in the panel regressions, with averages of 0.633 and 0.569 for DM and EM countries, respectively. The increase in *R-squared* when news variables are added to the baseline regression is quite large as well, averaging 0.161 (34%) for the DM sample and 0.149 (38%) for EM. Again, the added value varies strongly among countries. In the case of Nigeria, the increase is just 0.043, whereas Australia shows an increase of 0.455. From an econometric point of view it makes sense that the *R-squared* is higher in country-level OLS compared to panel regressions. By definition, panel regressions have one coefficient for each variable that is the same for all countries. In country-level OLS, these coefficients can be optimally fitted to the data of each country individually, leading to a higher *R-squared*.

In line with the 12-month returns, *R-squared* of monthly returns increased with country-level OLS compared to the panel regressions. For developed markets (emerging markets), the *R-squared* is 0.169 (0.213) on average and the increment over the baseline regression is 0.084 (0.086). In percentage terms this increment is larger than for 12-month returns, with 99% and 68% reported for DM and EM, respectively. Although the value of *R-squared* and the increment have improved, the average number of news variables that is significant is falling behind. Just 1.1 news variables were significant in developed markets, and 1.0 in emerging markets.

5.3 Results of Least Angle Regressions by country

For the evaluation of the importance of the news variables using the LARS procedure, regressions have been performed with s in equation (4.3) equal to 30%, 50%, and 70% of the sum of the absolute values of the coefficients in the unrestricted OLS regression. Subsequently, it has been determined what percentage of the news variables and what percentage of the other explanatory variables have non-zero coefficients. As these percentages were very similar for all values of s, I have only included the results of the 50% case in table 10 (12-month returns) and table 11 (monthly returns) for developed market countries. The results of the 50% case for EM countries are reported in the appendix. As is reported in Table 4, 8.1 and 8.4 (DM and EM, respectively) news variables are non-zero in this case, which is 68% and 70% of the total number of news variables. These numbers are only meaningful when compared to the same number for the other explanatory variables. Of these 10 return and macro variables, 84% and 88% remain non-zero, which is more than 15% higher than for news variables. For monthly returns the percentage difference of non-zero coefficients is even greater, 27% for DM and 20% for EM. Hence, the LARS method consistently eliminates more news variables than other variables, for each value of s and for each group of countries.

Table 10: Developed markets country-level Least Angle Regression for 12-month returns
Country-level Least Angle Regressions (LARS) for developed market countries of 12-month returns. The same variables have been used as in the panel regression reported in Table 5, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "0" denotes a coefficient that is set to zero in the regression. *Zeros* reports the total number of variables set to zero, and *Non-zeros* the number of variables that were not set to zero. *Nobs* reports the number of observations.

	AU	AT	BE	CA	DK	FI	FR	DE	GR	IE	IT	JP
$entropy_{t-1}$				0		0	0	0		0	0	
$artcount_{t-1}$			0	0		0				0		
$sMkt_{t-1}$									0		0	
$fMkt_{t-1}$		0				0	0	0		0		
$sGovt_{t-1}$			0		0	0	0				0	0
$fGovt_{t-1}$	0	0				0	0				0	0
$sCorp_{t-1}$	0	0		0	0			0	0	0	0	0
$fCorp_{t-1}$			0									
$sComms_{t-1}$	0	0	0	0	0	0	0	0				0
$fComms_{t-1}$				0						0	0	
$sCredit_{t-1}$	0	0			0	0	0			0		
$fCredit_{t-1}$	0	0	0	0	0	0	0	0	0	0	0	0
Zeros	5	6	5	6	5	8	7	5	3	7	7	5
Nonzeros	7	6	7	6	7	4	5	7	9	5	5	7
Nobs	211	179	161	139	175	179	179	179	176	179	134	211

	LU	NL	NZ	NO	PT	SG	ES	SE	CH	GB	US
$entropy_{t-1}$	0	0				0	0		0		
$artcount_{t-1}$	0	0					0				
$sMkt_{t-1}$	0				0	0			0		
$fMkt_{t-1}$		0	0	0	0		0			0	0
$sGovt_{t-1}$	0	0		0		0	0	0			
$fGovt_{t-1}$	0	0	0	0		0	0	0	0		0
$sCorp_{t-1}$			0	0		0			0	0	
$fCorp_{t-1}$		0			0						
$sComms_{t-1}$	0	0	0			0	0	0			
$fComms_{t-1}$		0			0	0			0	0	
$sCredit_{t-1}$	0	0			0	0	0				0
$fCredit_{t-1}$	0	0	0	0	0	0	0	0	0	0	0
Zeros	8	10	5	5	6	9	8	4	6	4	4
Non-zeros	4	2	7	7	6	3	4	8	6	8	8
Nobs	138	179	187	174	179	132	179	211	211	211	211

Table 11: Developed markets country-level Least Angle Regression for monthly returns
Country-level Least Angle Regressions (LARS) for developed market countries of monthly returns. The same variables have been used as in the panel regression reported in Table 6, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "0" denotes a coefficient that is set to zero in the regression. *Zeros* reports the total number of variables set to zero, and *Non-zeros* the number of variables that were not set to zero. Nobs reports the number of observations.

	AU	AT	BE	CA	DK	FI	FR	DE	GR	IE	IT	JP
$entropy_{t-1}$							0				0	
$artcount_{t-1}$		0										0
$sMkt_{t-1}$			0		0	0						
$fMkt_{t-1}$			0									
$sGovt_{t-1}$	0			0							0	
$fGovt_{t-1}$	0	0							0	0		
$sCorp_{t-1}$	0				0		0				0	0
$fCorp_{t-1}$		0				0	0	0		0	0	0
$sComms_{t-1}$		0						0				
$fComms_{t-1}$											0	
$sCredit_{t-1}$					0		0	0		0		
$fCredit_{t-1}$	0	0	0	0	0	0	0	0	0	0	0	0
Zeros	4	5	3	2	4	3	5	4	2	4	6	4
Non-zeros	8	7	9	10	8	9	7	8	10	8	6	8
Nobs	211	179	161	139	175	179	179	179	176	179	134	211

	LU	NL	NZ	NO	PT	SG	ES	SE	CH	GB	US
$entropy_{t-1}$			0			0		0		0	
$artcount_{t-1}$					0			0			
$sMkt_{t-1}$	0			0							
$fMkt_{t-1}$								0			
$sGovt_{t-1}$											
$fGovt_{t-1}$					0				0		
$sCorp_{t-1}$	0		0		0	0			0		
$fCorp_{t-1}$		0		0	0				0		
$sComms_{t-1}$											0
$fComms_{t-1}$		0			0		0		0	0	0
$sCredit_{t-1}$			0	0	0		0			0	
$fCredit_{t-1}$	0	0	0	0	0	0	0	0	0	0	0
Zeros	3	3	4	4	7	3	3	4	5	4	3
Non-zeros	9	9	8	8	5	9	9	8	7	8	9
Nobs	138	179	187	174	179	132	179	211	211	211	211

OLS regressions have been performed that only include the news variables with non-zero coefficients, and it turns out that the loss in *R-squared* is not large. Whereas including all news variables yielded *R-squared* increments of 34% for DM and 38% for EM in 12-month returns, including just the non-zero coefficient variables from LARS regressions yields *R-squared* increments of 30% and 33%. In other words, excluding 30% of the news variables barely had any impact on *R-squared*.

5.4 Results of WLS regressions by country

The results of the country-level WLS regressions for 12-month returns corrected for overlapping data issues are reported in Table 12 for developed markets and in the appendix for emerging markets. It becomes clear that corrected for the overlapping returns has taken away part of the (unjustified) predictive power of the variables, as the reported *R-squared* is far lower than in the country-level OLS regressions. The average *R-squared* is 0.241 for developed countries (versus 0.633 in OLS). However, the added value of the news variables in terms of *R-squared* increment is still present, marked by an average increase of 0.142 (143%). For emerging markets the results are similar but slightly less impressive, with an *R-squared* of 0.240 and an increment of 0.083 or 53%. Looking at the significance of individual news variables, the variation in results by country is striking. There are two countries for which 8 of the 12 news variables have a significant coefficient at the 5% level. At the same time, for 8 of the 23 countries in the developed markets sample none of the news variables is significant. Even though there is still an increase in *R-squared*, the fact that over 30% of the countries report no significant coefficients for the news variables when overlapping returns are corrected for strongly questions the predictive value of the authors' variables.

The results for the WLS regressions using U.S. data to predict returns in other countries are very surprising. The WLS regression tables (Table 12 for DM, appendix for EM) report the *R-squared* and *increment in %* when U.S. data is used. Both *R-squared* and *increment in %* are on average higher when U.S. data is used instead of country-individual data, which holds for DM as well as EM. Table 4 shows average *R-squared* for developed countries of 0.270 (EM: 0.262) when U.S. data is used, and *R-squared* increment of 143% (EM: 76%). This finding shows resemblance with the finding of Rapach et al. (2013), who showed that lagged U.S. returns were better predictors of future returns in developed countries than lagged returns of these countries themselves. However, the predictive value of U.S. variables also holds in the case of emerging market countries, which is something that was not investigated before.

Table 12: Developed markets country-level WLS for 12-month returns corrected for overlapping data issues
Country-level Weighted Least Squares (WLS) regressions for developed market countries of 12-month returns corrected for overlapping data issues. The same variables have been used as in the panel regression reported in Table 5, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "+" ("-") denotes a positive (negative) coefficient significant at the 5% level. *Total* reports the total number of significant news variables. "R2" and "R2 increment" denote the *R-squared* of the regression including news variables and the increment over the corresponding baseline (regression without news variables), respectively. The standard errors are obtained using the Newey-West procedure.

	AU	AT	BE	CA	DK	FI	FR	DE	GR	IE	IT	JP
$entropy_{t-1}$								-				
$artcount_{t-1}$		-										
$sMkt_{t-1}$		+								+		
$fMkt_{t-1}$		+										
$sGovt_{t-1}$	-							-				
$fGovt_{t-1}$		+										
$sCorp_{t-1}$												
$fCorp_{t-1}$		+										
$sComms_{t-1}$		+										
$fComms_{t-1}$		+										
$sCredit_{t-1}$								+				+
$fCredit_{t-1}$		+										
Total	1	8	0	0	0	0	0	3	0	1	0	1
R2	0.200	0.298	0.297	0.297	0.236	0.180	0.185	0.236	0.218	0.255	0.315	0.179
R2 increment	0.152	0.190	0.166	0.128	0.149	0.091	0.097	0.160	0.105	0.133	0.157	0.111
R2 - US	0.227	0.328	0.366	0.323	0.305	0.247	0.263	0.227	0.284	0.318	0.274	0.158
R2 increment - US	0.141	0.212	0.21	0.167	0.175	0.154	0.162	0.135	0.192	0.208	0.104	0.088
Nobs	211	179	161	139	175	179	179	179	176	179	134	211

	LU	NL	NZ	NO	PT	SG	ES	SE	CH	GB	US
$entropy_{t-1}$							+			-	
$artcount_{t-1}$			-						-	+	+
$sMkt_{t-1}$		+		+		+	+				
$fMkt_{t-1}$			+						-	+	
$sGovt_{t-1}$		-	-		-			-			
$fGovt_{t-1}$			+						-	+	
$sCorp_{t-1}$				-				+			
$fCorp_{t-1}$			+						-	+	
$sComms_{t-1}$				+		-		+			
$fComms_{t-1}$			+						-	+	
$sCredit_{t-1}$										+	
$fCredit_{t-1}$			+						-	+	
Total	0	2	7	3	1	2	2	3	6	8	1
R2	0.255	0.245	0.271	0.287	0.274	0.255	0.249	0.194	0.227	0.22	0.177
R2 increment	0.186	0.147	0.235	0.173	0.104	0.155	0.170	0.118	0.112	0.134	0.104
R2 - US	0.368	0.284	0.286	0.278	0.312	0.281	0.236	0.179	0.234	0.259	0.177
R2 increment - US	0.158	0.183	0.226	0.095	0.219	0.135	0.166	0.109	0.147	0.166	0.103
Nobs	138	179	187	174	179	132	179	211	211	211	211

6 Conclusion

This paper replicated the results of Calomiris and Mamaysky (2019), and subsequently evaluated the predictive value of news variables on country-level stock returns. I started with Ordinary Least Squares (OLS) regressions and Least Angle Regressions (LARS) for both 12-month returns and monthly returns. Weighted Least Squares (WLS) regressions that correct for overlapping data issues present in 12-month returns were then performed, which allowed for an econometrically correct reevaluation of the results obtained by Calomiris and Mamaysky (2019). The same WLS regressions have been performed for each country using U.S. data. I will now draw conclusions from the results by formulating answers to the three research questions defined in section 1.

1. *Are the news variables of Calomiris and Mamaysky (2019) still of predictive value when corrected for overlapping data issues?*

   (a) *Is the value of R-squared and the increment in R-squared when including news variables of similar magnitude in WLS regressions compared to the panel regressions of Calomiris and Mamaysky (2019)?*

   (b) *Do the coefficients of individual news variables remain significant in the WLS regressions?*

The answer to the first research question is yes, the news variables are of predictive value, even when corrected for overlapping data issues. The average increment in *R-squared* when including the news variables was over 50% for both EM and DM countries, which is in line or even better than the panel regressions. Furthermore, it is better than the results obtained from panel regressions of monthly returns. The latter result argues in favor of the authors' claim that monthly news variables show stronger predictive value for one-year ahead returns than for one-month ahead returns.

The number of significant news variables is similar in WLS and the panel regressions. However, still only 20% of the news variables are significant at the 5% level and adding all 12 news variables seems like an overkill. This is confirmed by the fact that *R-squared* barely dropped when excluding the news variables eliminated by the LARS procedure. To conclude, the model including 12 news variables has predictive value, but at the same time seems overly complex and could benefit from simplification.

2. *Are there large differences in forecasting value of news variables on a country level as indicated by country-level OLS, LARS and WLS regressions?*

Country-level regressions performed in this paper revealed a large variation in the predictive value of news variables from one country to the other, something that is unobservable in the aggregate panel regressions of Calomiris and Mamaysky (2019). In the WLS and OLS regressions of some countries, *R-squared* increased by less than 10% when news variables were included. Furthermore, there are countries for which none of the news variables is signifant at the 5% level. These are not just rare exceptions; for WLS regressions in developed markets this holds for 35% (8 of 23) of the countries. In the LARS regressions, there are 9 countries for which over half of the news variables has a zero coefficient. Therefore, it is highly relevant for investors to further investigate the predictive value of news variables in their countries of interest, instead of following the conclusions derived from panel regressions. The fact that the results of Calomiris and Mamaysky (2019) are not consistent across countries, negatively impacts the usefulness and reliability of their findings.

3. *Are the United States' news variables able to predict returns in other countries?*

The results of this paper show that U.S. news variables are on average better predictors of a country's future stock returns than a country's own news variables, indicated by a higher (increase in) *R-squared*. This can be explained by the theory of gradual information diffusion, which states that new information is first incorporated in the price of broadly covered (U.S.) assets and incorporated with a delay in the prices of other assets. This research has found that this does not only hold for developed countries, but also for emerging market countries.

The findings of this paper open the door for two areas of potential future research. The insignificance of many news variables in OLS and WLS regressions and their elimination in LARS procedures suggest that the model employed by Calomiris and Mamaysky (2019) should be simplified. One way to do this is by dropping topic-specific variables and focus on general sentiment and article frequency instead, which is in line with earlier research by for example Tetlock (2007) and Garcia (2013). This would decrease the number of news variables from 12 to just 3: *entropy*, *article frequency* and *sentiment*. I expect that this change will also eliminate the collinearity issue and improve the significance of the coefficients of the individual variables.

Another area of future research could built on the discovered predictive value of U.S. news for stock markets worldwide. Since Rapach et al. (2013) argue that the predictive value of U.S. returns is bigger for countries that are major trade partners, it would be interesting to see if this relationship holds for the predictive value of U.S. news as well. In other words, examine whether there exists a correlation between the predictive value of U.S. news for a country's stock returns and the relative size of the trade activity between the U.S. and this country.

7 Appendix

Table A1: Emerging markets forecasting panel for 12-month returns

Panel regressions for emerging market 12-month returns ($return^{12}$). Significant coefficients at the 1%, 5%, and 10% level are labeled with "***", "**", and "*", respectively. Descriptions of the variables in the regressions can be found in Table 2. The *Base*, *Sent*, and *SentEnt* columns report the baseline regression, sentiment regression, and sentiment-entropy regression for the entire period and two subperiods. The *stderr* row indicates how the standard errors are clustered, where *both* means double-clustered by country and time. The absolute values of the coefficients of the news variables are larger in the replication than in the original paper. This is due to the fact that this paper standardizes the news variables by subtracting the mean and dividing by the standard deviation, whereas the authors did not subtract the mean.

	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	0.235*	0.364***	0.350***	0.129	0.409***	0.402***	0.474***	0.492***	0.473***
$sigma_{t-2}$	0.344**	0.371***	0.360***	0.177	0.301**	0.300**	0.508***	0.519***	0.503***
$return_{t-1}$	0.203	0.061	0.081	-0.011	-0.221	-0.208	0.329	0.234	0.257
$return_{t-2}$	0.005	-0.060	-0.049	-0.069	-0.149	-0.142	-0.023	-0.045	-0.033
$value_{t-1}$	0.508	4.771	4.680	1.114	9.391**	9.167**	4.824	1.028	1.247
gdp_{t-1}	-1.443*	-1.603**	-1.555**	-0.300	-1.435	-1.438	-3.516***	-3.207***	-3.116***
$gdpdeflator_{t-1}$	0.329	0.167	0.179	0.639	0.010	0.012	-0.623	-0.532	-0.537
cp_{t-1}	-0.683**	-0.394*	-0.400*	-0.302	-0.103	-0.106	0.014	-0.050	-0.066
$rate_{t-1}$	-0.488	-0.399	-0.398	-0.588	0.198	0.204	-1.616	-1.640	-1.557
$dexch_{t-1}$	0.027	0.113	0.125	-0.273	-0.151	-0.145	0.413	0.407	0.400
$entropy_{t-1}$		-25.678	-24.395		-79.260*	-77.651*		30.519	28.163
$artcount_{t-1}$		-13.394*	-13.037*		-41.204**	-40.561**		-9.492	-9.021
$sMkt_{t-1}$		7.827**	6.897*		4.904	3.450		8.835**	7.802**
$fMkt_{t-1}$		-77.276**	-72.110**		-43.930	-43.697		-3.050	0.789
$sGovt_{t-1}$		-0.148	1.629		10.055**	10.926***		-5.391	-3.813
$fGovt_{t-1}$		-88.825**	-80.819**		-36.108	-33.979		-20.478	-14.379
$sCorp_{t-1}$		-9.778*	-10.171*		-4.377	-4.104		-10.478*	-10.772*
$fCorp_{t-1}$		-38.138***	-35.815**		-18.906	-18.170		-4.117	-2.375
$sComms_{t-1}$		-0.777	0.722		0.273	1.501		0.566	1.475
$fComms_{t-1}$		-49.584**	-44.059*		-7.480	-5.878		-7.968	-3.875
$sMacro_{t-1}$		7.668	5.443		1.804	1.677		-3.034	-5.203
$fMacro_{t-1}$		omitted	omitted		omitted	omitted		omitted	omitted
R2	0.1287	0.2178	0.2145	0.1603	0.3141	0.3148	0.3535	0.3928	0.3931
start	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4365	4365	4365	2013	2013	2013	2352	2352	2352
stderr	both	both	both	both	both	both	both	both	both

Table A2: Emerging markets forecasting panel for monthly returns
Panel regressions for developed market monthly returns (*return*). Significant coefficients at the 1%, 5%, and 10% level are labeled with "***", "**", and "*", respectively. Descriptions of the variables in the regressions can be found in Table 2. The *Base*, *Sent*, and *SentEnt* columns report the baseline regression, sentiment regression, and sentiment-entropy regression for the entire period and two subperiods. The *stderr* row indicates how the standard errors are clustered, where *both* means double-clustered by country and time. The absolute values of the coefficients of the news variables are larger in the replication than in the original paper. This is due to the fact that this paper standardizes the news variables by subtracting the mean and dividing by the standard deviation, whereas the authors did not subtract the mean.

	Base	Sent	SentEnt	Base	Sent	SentEnt	Base	Sent	SentEnt
$sigma_{t-1}$	-0.030	-0.023	-0.024	-0.000	0.02	0.019	0.448***	-0.044	-0.045
$sigma_{t-2}$	0.040	0.042	0.042	0.002	0.011	0.011	0.597***	0.081	0.081
$return_{t-1}$	0.104**	0.087*	0.090*	0.057	0.038	0.04	0.382	0.099	0.102
$return_{t-2}$	0.002	-0.004	-0.002	-0.045	-0.052	-0.051	0.085	0.026	0.028
$value_{t-1}$	0.116	0.165	0.173	0.314	1.045**	1.020**	13.065*	-0.305	-0.293
gdp_{t-1}	-0.073	-0.077	-0.080	-0.002	-0.116	-0.116	-2.731***	-0.171	-0.174
$gdpdeflator_{t-1}$	-0.074	-0.081	-0.081	-0.032	-0.089	-0.088	-0.728**	-0.241**	-0.245**
cp_{t-1}	-0.047**	-0.034*	-0.034*	-0.027**	-0.025	-0.025	-0.081	-0.047	-0.048
$rate_{t-1}$	0.061	0.054	0.055	0.054	0.108*	0.109*	-1.43	-0.305	-0.304
$dexch_{t-1}$	-0.006	-0.004	-0.005	-0.018	-0.010	-0.010	0.623	0.022	0.020
$entropy_{t-1}$		2.676	2.842		-8.492**	-8.112*		8.545***	8.858***
$artcount_{t-1}$		-0.170	-0.164		-1.060	-1.037		-0.352	-0.341
$sMkt_{t-1}$		0.718	0.558		0.166	0.026		0.709	0.534
$fMkt_{t-1}$		-1.761*	-1.937*		-2.953	-3.107		-1.359	-1.744
$sGovt_{t-1}$		-0.041	-0.027		-0.019	0.017		0.373	0.273
$fGovt_{t-1}$		-2.480**	-2.568**		-2.043	-2.093		-2.500	-2.941*
$sCorp_{t-1}$		-0.816*	-0.709		0.163	0.137		-1.091***	-0.953**
$fCorp_{t-1}$		-1.440***	-1.425***		-0.921	-0.931		-1.386**	-1.462**
$sComms_{t-1}$		0.829**	0.837**		0.516	0.627		1.165**	1.213**
$fComms_{t-1}$		-0.519	-0.635		-0.116	-0.109		-0.966	-1.277
$sMacro_{t-1}$		-0.215	-0.161		0.182	0.259		-0.643	-0.556
$fMacro_{t-1}$		omitted	omitted		omitted	omitted		omitted	omitted
R2	0.029	0.039	0.038	0.020	0.039	0.039	0.054	0.084	0.083
start	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Apr 1998	Mar 2007	Mar 2007	Mar 2007
end	Dec 2015	Dec 2015	Dec 2015	Feb 2007	Feb 2007	Feb 2007	Dec 2015	Dec 2015	Dec 2015
Nobs	4365	4365	4365	2013	2013	2013	2352	2352	2352
stderr	both	both	both	both	both	both	both	both	both

Table A3: Emerging markets country-level OLS for 12-month returns

Country-level Ordinary Least Squares (OLS) regressions for emerging market countries of next 12-month returns ($return^{12}$). The same variables have been used as in the panel regression reported in Table A1, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "+" ("-") denotes a positive (negative) coefficient significant at the 5% level. *Total* reports the total number of significant news variables. "R2" and "R2 increment" denote the *R-squared* of the regression including news variables and the increment over the corresponding baseline (regression without news variables), respectively. The standard errors are obtained using the Newey-West procedure (Newey and West, 1987).

	AR	BR	CL	CN	CO	CZ	EE	GH	HK	HU	IN	ID	IL	KE
$entropy_{t-1}$		-								+	+		-	
$artcount_{t-1}$	-		-	+						-	-	-		
$sMkt_{t-1}$	+								+			+		
$fMkt_{t-1}$		+			+						+		-	
$sGovt_{t-1}$		-										+		
$fGovt_{t-1}$														
$sCorp_{t-1}$		+				-			-					
$fCorp_{t-1}$						-								
$sComms_{t-1}$	-			+										
$fComms_{t-1}$				+					+		+			
$sMacro_{t-1}$		+					+	-		-		-		
$fMacro_{t-1}$		+	+				+					-		
Total	3	6	2	3	1	2	2	1	3	3	4	5	2	0
R2	0.600	0.538	0.394	0.512	0.687	0.726	0.709	0.655	0.519	0.509	0.615	0.497	0.365	0.696
R2 increment	0.119	0.246	0.066	0.251	0.111	0.077	0.193	0.111	0.260	0.256	0.180	0.166	0.104	0.129
Nobs	186	211	211	211	159	195	143	129	211	205	165	211	211	134

	MY	MX	NG	PE	PH	PL	RU	SK	SQ	ZA	KR	TH	TR	UA
$entropy_{t-1}$	-					-			+				-	-
$artcount_{t-1}$				-	-				-	-	+			-
$sMkt_{t-1}$	+							-	+	+		+		
$fMkt_{t-1}$				+							+	+	-	
$sGovt_{t-1}$		-							-		-		-	
$fGovt_{t-1}$									-					
$sCorp_{t-1}$				-					-		+	-		
$fCorp_{t-1}$		+				-			-					
$sComms_{t-1}$							-							
$fComms_{t-1}$		+		+	-			+			+		+	
$sMacro_{t-1}$				+						-			+	+
$fMacro_{t-1}$	+			+		+							+	+
Total	3	3	0	6	2	3	1	2	7	3	5	3	6	4
R2	0.588	0.392	0.547	0.628	0.463	0.618	0.649	0.719	0.382	0.569	0.777	0.429	0.406	0.735
R2 increment	0.136	0.151	0.043	0.169	0.229	0.099	0.050	0.172	0.104	0.226	0.101	0.149	0.157	0.119
Nobs	211	211	140	211	211	195	155	119	143	183	211	154	211	179

Table A4: Emerging markets country-level OLS for monthly returns
Country-level Ordinary Least Squares (OLS) regressions for emerging market countries of monthly returns. The same variables have been used as in the panel regression reported in Table A2, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "+" ("-") denotes a positive (negative) coefficient significant at the 5% level. *Total* reports the total number of significant news variables. "R2" and "R2 increment" denote the *R-squared* of the regression including news variables and the increment over the corresponding baseline (regression without news variables), respectively. The standard errors are obtained using the Newey-West procedure (Newey and West, 1987).

	AR	BR	CL	CN	CO	CZ	EE	GH	HK	HU	IN	ID	IL	KE
$entropy_{t-1}$														
$artcount_{t-1}$					+									
$sMkt_{t-1}$		-				+								
$fMkt_{t-1}$														
$sGovt_{t-1}$													+	
$fGovt_{t-1}$								+						
$sCorp_{t-1}$	-					-								
$fCorp_{t-1}$														
$sComms_{t-1}$						+	-			+				+
$fComms_{t-1}$														
$sMacro_{t-1}$		+												
$fMacro_{t-1}$		+				+	+							
Total	1	3	0	0	1	4	2	1	0	1	0	0	1	1
R2	0.147	0.142	0.160	0.152	0.244	0.208	0.255	0.334	0.091	0.246	0.132	0.133	0.097	0.406
R2 increment	0.072	0.088	0.028	0.067	0.030	0.113	0.177	0.169	0.043	0.156	0.067	0.045	0.041	0.091
Nobs	186	211	211	211	159	195	143	129	211	205	165	211	211	134

	MY	MX	NG	PE	PH	PL	RU	SK	SQ	ZA	KR	TH	TR	UA
$entropy_{t-1}$										+				
$artcount_{t-1}$					-									
$sMkt_{t-1}$														
$fMkt_{t-1}$										+		+		
$sGovt_{t-1}$														
$fGovt_{t-1}$														
$sCorp_{t-1}$						-			-		+			
$fCorp_{t-1}$						-					+			
$sComms_{t-1}$								+						
$fComms_{t-1}$													+	
$sMacro_{t-1}$														
$fMacro_{t-1}$				+										
Total	0	0	0	1	1	2	0	1	1	2	2	1	1	0
R2	0.239	0.155	0.275	0.310	0.127	0.154	0.259	0.250	0.339	0.134	0.235	0.203	0.134	0.211
R2 increment	0.091	0.095	0.029	0.052	0.070	0.100	0.047	0.153	0.181	0.095	0.106	0.050	0.047	0.033
Nobs	211	211	140	211	211	195	155	119	143	183	211	154	211	179

Table A5: Emerging markets country-level Least Angle Regression for 12-month returns
Country-level Least Angle Regressions (LARS) for emerging market countries of 12-month returns. The same variables have been used as in the panel regression reported in Table A1, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "0" denotes a coefficient that is set to zero in the regression. *Zeros* reports the total number of variables set to zero, and *Non-zeros* the number of variables that were not set to zero. Nobs reports the number of observations.

	AR	BR	CL	CN	CO	CZ	EE	GH	HK	HU	IN	ID	IL	KE
$entropy_{t-1}$						0	0		0					
$artcount_{t-1}$					0	0	0	0	0					
$sMkt_{t-1}$		0	0	0						0				
$fMkt_{t-1}$	0			0	0			0	0					0
$sGovt_{t-1}$	0		0	0		0		0	0	0				
$fGovt_{t-1}$										0				
$sCorp_{t-1}$	0							0		0			0	
$fCorp_{t-1}$			0						0		0		0	0
$sComms_{t-1}$					0	0	0	0	0	0				
$fComms_{t-1}$	0				0	0		0		0			0	0
$sMacro_{t-1}$	0				0	0	0		0					
$fMacro_{t-1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zeros	6	2	4	4	6	7	5	7	8	7	2	1	4	4
Non-zeros	6	10	8	8	6	5	7	5	4	5	10	11	8	8
Nobs	186	211	211	211	159	195	143	129	211	205	165	211	211	134

	MY	MX	NG	PE	PH	PL	RU	SK	SQ	ZA	KR	TH	TR	UA
$entropy_{t-1}$						0	0							
$artcount_{t-1}$					0		0							
$sMkt_{t-1}$		0				0				0				
$fMkt_{t-1}$	0				0									
$sGovt_{t-1}$			0	0	0	0	0					0		
$fGovt_{t-1}$	0			0	0	0		0	0		0			
$sCorp_{t-1}$				0	0	0	0		0					
$fCorp_{t-1}$												0		
$sComms_{t-1}$	0		0	0	0		0	0						
$fComms_{t-1}$					0	0				0				0
$sMacro_{t-1}$							0			0				
$fMacro_{t-1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zeros	4	2	3	5	8	7	7	3	3	4	2	3	1	2
Non-zeros	8	10	9	7	4	5	5	9	9	8	10	9	11	10
Nobs	211	211	140	211	211	195	155	119	143	183	211	154	211	179

Table A6: Emerging markets country-level Least Angle Regression for monthly returns
Country-level Least Angle Regressions (LARS) for emerging market countries of monthly returns. The same variables have been used as in the panel regression reported in Table A2, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "0" denotes a coefficient that is set to zero in the regression. *Zeros* reports the total number of variables set to zero, and *Non-zeros* the number of variables that were not set to zero. *Nobs* reports the number of observations.

	AR	BR	CL	CN	CO	CZ	EE	GH	HK	HU	IN	ID	IL	KE
$entropy_{t-1}$											0			0
$artcount_{t-1}$							0				0			
$sMkt_{t-1}$													0	0
$fMkt_{t-1}$	0		0	0						0	0			0
$sGovt_{t-1}$		0	0	0		0	0							
$fGovt_{t-1}$										0				
$sCorp_{t-1}$		0								0			0	
$fCorp_{t-1}$	0													0
$sComms_{t-1}$											0			
$fComms_{t-1}$	0					0		0		0				
$sMacro_{t-1}$	0					0			0					0
$fMacro_{t-1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zeros	5	3	3	3	1	4	3	2	2	5	5	1	3	6
Non-zeros	7	9	9	9	11	8	9	10	10	7	7	11	9	6
Nobs	186	211	211	211	159	195	143	129	211	205	165	211	211	134

	MY	MX	NG	PE	PH	PL	RU	SK	SQ	ZA	KR	TH	TR	UA
$entropy_{t-1}$													0	
$artcount_{t-1}$	0											0	0	0
$sMkt_{t-1}$		0								0	0	0	0	
$fMkt_{t-1}$	0							0	0					
$sGovt_{t-1}$			0	0		0		0			0	0	0	
$fGovt_{t-1}$	0						0						0	0
$sCorp_{t-1}$				0	0		0			0			0	
$fCorp_{t-1}$										0				
$sComms_{t-1}$	0		0				0			0				
$fComms_{t-1}$		0									0	0		
$sMacro_{t-1}$		0		0								0	0	
$fMacro_{t-1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zeros	5	4	3	4	2	2	4	3	2	5	4	6	8	3
Non-zeros	7	8	9	8	10	10	8	9	10	7	8	6	4	9
Nobs	211	211	140	211	211	195	155	119	143	183	211	154	211	179

Table A7: Emerging markets country-level WLS for 12-month returns corrected for overlapping data issues
Country-level Weighted Least Squares (WLS) regressions for emerging market countries of 12-month returns corrected for overlapping data issues. The same variables have been used as in the panel regression reported in Table A1, but only the news variables are reported. Descriptions of all variables are provided in Table 2. The two-letter codes are defined by Thomson Reuters (TR codes) and a list of corresponding countries can be found in Table 1. A "+" ("-") denotes a positive (negative) coefficient significant at the 5% level. *Total* reports the total number of significant news variables. "R2" and "R2 increment" denote the *R-squared* of the regression including news variables and the increment over the corresponding baseline (regression without news variables), respectively. The standard errors are obtained using the Newey-West procedure.

	AR	BR	CL	CN	CO	CZ	EE	GH	HK	HU	IN	ID	IL	KE
$entropy_{t-1}$	+													-
$artcount_{t-1}$	+									-				
$sMkt_{t-1}$	+			+		+			+	+				
$fMkt_{t-1}$				-							+			+
$sGovt_{t-1}$		-								-				
$fGovt_{t-1}$				-							+			+
$sCorp_{t-1}$	-	+	+			-		+				+		-
$fCorp_{t-1}$				-							+			+
$sComms_{t-1}$				+			-							
$fComms_{t-1}$				-							+			+
$sMacro_{t-1}$				-				-				-		+
$fMacro_{t-1}$				-							+			+
Total	4	2	1	8	0	2	1	2	1	3	5	2	0	8
R2	0.260	0.172	0.23	0.246	0.241	0.214	0.354	0.268	0.127	0.204	0.260	0.195	0.136	0.368
R2 increment	0.072	0.088	0.028	0.067	0.030	0.113	0.177	0.169	0.043	0.156	0.067	0.045	0.041	0.091
R2 - US	0.178	0.188	0.216	0.254	0.25	0.248	0.347	0.709	0.17	0.177	0.269	0.176	0.165	0.398
R2 increment - US	0.072	0.088	0.028	0.067	0.03	0.113	0.177	0.169	0.043	0.156	0.067	0.045	0.041	0.091
Nobs	186	211	211	211	159	195	143	129	211	205	165	211	211	134

	MY	MX	NG	PE	PH	PL	RU	SK	SQ	ZA	KR	TH	TR	UA
$entropy_{t-1}$														
$artcount_{t-1}$			+	-		-					+			
$sMkt_{t-1}$	+	+						-	+		-			
$fMkt_{t-1}$			+	-						-			-	
$sGovt_{t-1}$					+				+	-			-	
$fGovt_{t-1}$			+	-						-			-	
$sCorp_{t-1}$									-		+			
$fCorp_{t-1}$			+	-						-			-	
$sComms_{t-1}$						+					+			
$fComms_{t-1}$			+	-						-			-	
$sMacro_{t-1}$				+										
$fMacro_{t-1}$			+	-						-			-	
Total	1	1	6	7	1	2	0	1	3	6	4	0	6	0
R2	0.178	0.122	0.359	0.234	0.182	0.221	0.279	0.302	0.287	0.194	0.171	0.397	0.159	0.355
R2 increment	0.091	0.095	0.029	0.052	0.070	0.100	0.047	0.153	0.181	0.095	0.106	0.050	0.047	0.033
R2 - US	0.229	0.187	0.293	0.267	0.226	0.208	0.278	0.298	0.413	0.234	0.172	0.299	0.114	0.377
R2 increment - US	0.091	0.095	0.029	0.052	0.07	0.1	0.047	0.153	0.181	0.095	0.106	0.05	0.047	0.033
Nobs	211	211	140	211	211	195	155	119	143	183	211	154	211	179

7.1 Programming Code

Below the programming code to run panel regressions, Least Angle Regressions and Ordinary Least Squares regressions has been included. The Least Angle Regressions code makes use of a method called *lars* written by Efron et al. (2004), which has been made available in a zip-file. The Weighted Least Squares code has been obtained from Johnson (2018) and has been included in a zip-file as well.

```
// Initialize the Panel Data

xtset CountryID Period

// Create table with descriptive statistics

tabstat Return Return12 Sigma Value GDP GDPDeflator CP Rate dexch entropy ///
artcount sMkt fMkt sGovt fGovt sCorp fCorp sComms fComms sCredit fCredit, ///
stat(mean sd p5 p95 n) col(stat)

// Code to get the AR(1) coefficients for Return, for other variables the code
// is identical.

tempname ARRegReturn
postfile `ARRegReturn' CountryID ARCoef using ARRegReturn.dta, replace
quietly forval x = 1/23 {
    reg Return L.Return if CountryID == `x'
    post `ARRegReturn' (`x') (`=_b[L.Return]')
}
postclose `ARRegReturn'

use ARRegReturn, clear
export excel ARCoef using "H:\Documents\JasperBal\Bsc thesis\ARCoeff.xlsx", ///
sheetmodify cell(A1)
```

Figure 2: *Stata* code to initialize the panel data and obtain the descriptive statistics of Table 3.

```
// Standardize the news variables

egen artcountZ = std(artcount)
egen sMktZ = std(sMkt)
egen fMktZ = std(fMkt)
egen sGovtZ = std(sGovt)
egen fGovtZ = std(fGovt)
egen sCorpZ = std(sCorp)
egen fCorpZ = std(fCorp)
egen sCommsZ = std(sComms)
egen fCommsZ = std(fComms)
...
egen esCommsZ = std(esComms)
egen esCreditZ = std(esCredit)

// Sent Panel regression with double-clustered standard errors by country
// and time period. This is the entire period developed markets panel
// regression, others work similar.

ssc install reghdfe //Installs the software from Correia (2016)
reghdfe, compile
reghdfe Return12 L.Sigma L2.Sigma L.Return L2.Return L.Value L.GDP ///
L.GDPDeflator L.CP L.Rate L.dexch L.entropy L.artcountZ l.sMktZ l.fMktZ ///
l.sGovtZ l.fGovtZ l.sCorpZ l.fCorpZ l.sCommsZ l.fCommsZ l.sCreditZ ///
l.fCreditZ, absorb(CountryID) cluster(CountryID Period)

// Regressions to examine the behavior of residuals (identify autocorrelation)

xtreg Return12 L.Sigma L2.Sigma L.Return L2.Return L. Value L.GDP ///
L.GDPDeflator L.CP L.Rate L.dexch L.entropy L.artcountZ l.sMktZ l.fMktZ ///
l.sGovtZ l.fGovtZ l.sCorpZ l.fCorpZ l.sCommsZ l.fCommsZ l.sCreditZ ///
l.fCreditZ, fe
predict e, residuals
tabulate CountryID, summarize(e) //residuals per country
tabulate Period, summarize(e) //residuals by time

// Regression to identify collinearity in developed markets case

xtreg fCredit Return Sigma Value GDP GDPDeflator CP Rate dexch entropy ///
artcount sMkt fMkt sGovt fGovt sCorp fCorp sComms fComms sCredit
```

Figure 3: *Stata* code to standardize the news variables, run the panel regressions with double-clustered standard errors, group residuals by country and time, and identify collinearity.

```
function [coeffs] = CountryLARS(AllData)

%AllData contains Returns, Sigma, other macro variables and news variables.
%coeffs are the coefficients of each variable in the LARS algorithm.

coeffs = zeros(22,1); %to this variable the results for each country are added

%every i corresponds to one country
for i = 1:23

%extract the data corresponding to country i from the whole database.
elementsbefore = sum(AllData(:,1) < i);
elements = sum(AllData(:,1) == i);
CountryData = AllData((elementsbefore + 1):(elementsbefore + elements), :);

%extract the return data and explanatory variables from the country data.
Return = CountryData(:, 5);
Return12 = CountryData(:, 6);
Sigma = CountryData(:,7);
otherMacro = CountryData(:, 8:13);
newsVar = CountryData(:, 14:25);

X = [Sigma(2:end-1) Sigma(1:end-2) Return(2:end-1) Return(1:end-2)...
        otherMacro(2:end-1, :) newsVar(2:end-1, :)];
Y = Return12(3:end);

%run the LARS algorithm with no restrictions on sum of coeffs (regular OLS)
beta  = lars(X, Y, 'lasso', Inf, 1); %LARS algorithm of Efron et al. (2004)

%set restriction to 50% of sum of coeffs
dimB = size(beta);
beta = beta(dimB(1,1),:);
totalcoeff = sum(abs(beta));
restriction = 0.5* totalcoeff;

%run LARS with the restriction and add to variable coeffs
beta  = lars(X, Y, 'lasso', restriction, 1);
beta = transpose(beta);
dimBeta = size(beta);
NoCol = dimBeta(1,2);
coeffs = [coeffs beta(:,NoCol)];

end
```

Figure 4: *Matlab* code to run the Least Angle Regressions.

```
tempname OLSCountry

postfile `OLSCountry' CountryID CSig SESig CSig2 SESig2 CRet SERet CRet2 SERet2 CVal ///
SEVal CGDP SEGDP CDefl SEDefl CCP SECP CRate SERate Cdexch SEdexch Cent SEent Cart SEart ///
Cfmkt SEfmkt Csmkt SEsmkt Csgov SEsgov Cfgov SEfgov Cscorp SEscorp Cfcorp SEfcorp Cscomms SEscomms Cfcomms ///
SEfcomms Cscred SEscred Cfcred SEfcred R2 Obs using OLSCountry.dta, replace

// every x corresponds to one counntry

quietly forval x = 1/23 {

    // run the regression
    reg Return12 L.Sigma L2.Sigma L.Return L2.Return L.Value L.GDP ///
    L.GDPDeflator L.CP L.Rate L.dexch L.entropy L.artcountZ l.sMktZ l.fMktZ ///
    l.sGovtZ l.fGovtZ l.sCorpZ l.fCorpZ l.sCommsZ l.fCommsZ l.sCreditZ ///
    l.fCreditZ if CountryID == `x'

    //save the coefficients and its standard error
    post `OLSCountry' (`x') (`=_b[L.Sigma]') (`=_se[L.Sigma]') (`=_b[L2.Sigma]') (`=_se[L2.Sigma]') (`=_b[L.Return]') ///
    (`=_se[L.Return]')  (`=_b[L2.Return]') (`=_se[L2.Return]') (`=_b[L.Value]') (`=_se[L.Value]') (`=_b[L.GDP]') ///
    (`=_se[L.GDP]')(`=_b[L.GDPDeflator]') (`=_se[L.GDPDeflator]') (`=_b[L.CP]') (`=_se[L.CP]') (`=_b[L.Rate]') ///
    (`=_se[L.Rate]') (`=_b[L.dexch]') (`=_se[L.dexch]') ///
    (`=_b[L.entropy]') (`=_se[L.entropy]') (`=_b[L.artcountZ]') (`=_se[L.artcountZ]') ///
    (`=_b[l.sMktZ]') (`=_se[l.sMktZ]') (`=_b[l.fMktZ]') (`=_se[l.fMktZ]') (`=_b[l.sGovtZ]') (`=_se[l.sGovtZ]') ///
    (`=_b[l.fGovtZ]') (`=_se[l.fGovtZ]') (`=_b[l.sCorpZ]') (`=_se[l.sCorpZ]') (`=_b[l.fCorpZ]') (`=_se[l.fCorpZ]')
    (`=_b[l.sCommsZ]') (`=_se[l.sCommsZ]') (`=_b[l.fCommsZ]') (`=_se[l.fCommsZ]') (`=_b[l.sCreditZ]')
    (`=_se[l.sCreditZ]') (`=_b[l.fCreditZ]') (`=_se[l.fCreditZ]') ///
    (`=e(r2)') (`=(e(N))')
}
postclose `OLSCountry'

//export all results to excel
use OLSCountry, clear
export excel CountryID CSig SESig CSig2 SESig2 CRet SERet CRet2 SERet2 CVal ///
SEVal CGDP SEGDP CDefl SEDefl CCP SECP CRate SERate Cdexch SEdexch Cent SEent Cart SEart ///
Cfmkt SEfmkt Csmkt SEsmkt Csgov SEsgov Cfgov SEfgov Cscorp SEscorp Cfcorp SEfcorp Cscomms SEscomms Cfcomms ///
SEfcomms Cscred SEscred Cfcred SEfcred R2 Obs using "H:\Documents\JasperBal\Bsc thesis\back on drive\OLSCountry.xlsx", ///
sheetmodify cell(A2)

```

Figure 5: *Stata* code to run the Ordinary Least Squares regressions.

References

[1] Basher, S. A., Haug, A. A., & Sadorsky, P. (2012). Oil prices, exchange rates and emerging stock markets. *Energy Economics, 34(1)*, 227-240.

[2] Calomiris, C. W., Love, I., Pería, M. S. M. (2012). Stock returns' sensitivities to crisis shocks: Evidence from developed and emerging markets. *Journal of International Money and Finance, 31(4)*, 743-765.

[3] Calomiris, C. W., & Mamaysky, H. (2019). How news and its context drive risk and returns around the world. *Journal of Financial Economics.*

[4] Charest, G. (1978). Dividend information, stock returns and market efficiency-II. *Journal of Financial Economics, 6(2-3)*, 297-330.

[5] Correia, S. (2016). A feasible estimator for linear models with multi-way fixed effects. *Duke University Preliminary Version.* URL: www.scorreia.com/research/hdfe.pdf.

[6] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics, 32(2)*, 407-499.

[7] Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. *CRC press.*

[8] Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International economic review, 10(1)*, 1-21.

[9] Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy, 81(3)*, 607-636.

[10] Fama, E. F., & Schwert, G. W. (1977). Asset returns and inflation. *Journal of financial economics, 5(2)*, 115-146.

[11] Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance, 68(3)*, 1267-1300.

[12] Greene, W. H. (2003). Econometric analysis. *Pearson Education India.*

[13] Hansen, L. P., Hodrick, R. J. (1980). Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of political economy, 88(5)*, 829-853.

[14] Harri, A., & Brorsen, B. W. (1998). The overlapping data problem. *Available at SSRN 76460.*

[15] Heston, S. L., & Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal, 73(3)*, 67-83.

[16] Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *The Review of Financial Studies, 5(3)*, 357-386.

[17] Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. *Journal of financial economics, 6(2/3)*, 95-101.

[18] Johnson, T. L. (2018). A fresh look at return predictability using a more efficient estimator. *Review of Asset Pricing Studies*, Forthcoming.

[19] La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. W. (1998). Law and finance. *Journal of political economy, 106(6)*, 1113-1155.

[20] Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777-787.

[21] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences, 103(23)* , 8577-8582.

[22] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E, 69(2)*, 026113.

[23] Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies, 22(1)*, 435-480.

[24] Rapach, D. E., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies, 23(2)*, 821-862.

[25] Rapach, D. E., Strauss, J. K., & Zhou, G. (2013). International stock return predictability: what is the role of the United States? *The Journal of Finance, 68(4)*, 1633-1662.

[26] Rizova, S., (2010). Predictable trade ows and returns of trade-linked countries. *In AFA 2011 Denver Meetings Paper*

[27] Rogers, W. (1994). Regression standard errors in clustered samples. *Stata technical bulletin,* 3(13).

[28] Sinha, N. R. (2016). Underreaction to news in the US stock market. *Quarterly Journal of Finance, 6(02)* , 1650005.

[29] Thorbecke, W. (1997). On stock market returns and monetary policy. *The Journal of Finance, 52(2)*, 635-654.

[30] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance, 62(3)*, 1139-1168.

[31] Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63(3)*, 1437-1467.

[32] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58(1)*, 267-288.

[33] Welch, I., & Goyal, A. (2007). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies, 21(4)*, 1455-1508.