Sensitivity to Sentiment: News vs Social Media*

Baoqing Gan[♯,†], Vitali Alexeev[♯,‡], Ron Bird[♮], Danny Yeung[♯,]

♯ Finance Department, UTS Business School, University of Technology Sydney
Sydney, New South Wales 2007, Australia
‡ Department of Economics and Finance, University of Guelph, Ontario N1G2W1, Canada
♮ Waikato Management School, University of Waikato, Hamilton 3240, New Zealand

September 13, 2019

Abstract

We explore the rapidly changing social and news media landscape that is responsible for the dissemination of information vital to the efficient functioning of the financial markets. Using the sheer volume of social and news media activity, commonly known as buzz, we document three distinct regimes. We find that between 2011 and 2013 the news media coverage stimulates activity in social media. This is followed by a transition period of two-way causality. From 2016, however, changes in levels of social media activity seem to lead and generate news coverage volumes. We uncover similar evolution of lead-lag pattern between sentiment measures constructed from the tonality contained in textual data from social and news media posts. We discover that market variables exert stronger impact on investor sentiment than the other way around. We also find that return responses to social media sentiment almost doubled after the transition period, while return responses to news-based sentiment almost halved to its pre-transition level. The linkage between volatility and sentiment is much more persistent than that between returns and sentiment. Overall, our results suggest that social media is becoming the dominant media source.

Keywords: investor sentiment; textual analysis; vector autoregressive (VAR) model; TRMI
JEL: G14, G40, G41

*We thank Tianyu Zhang and participants at the WRDS 2019 Advanced Research Scholar Programme for insightful suggestions. We thank Terry Walter, Christina Nikitopoulos Sklibosios and other participants' helpful comments at the 8th SIRCA Young Researchers Workshop. We are grateful to Roberto Pascual, Shan Chen and other participants' comments from 2018 New Zealand Finance Meeting. We appreciate to Romain Legrand, Alistair Dieppe, and Björn van Roye from European Central Bank (ECB) for providing MATLAB's **Bayesian Estimation, Analysis and Regression (BEAR) Toolbox**; we thank Thomson Reuters Financial and Risk for offering MarketPsych Indices (**TRMI**) as part of our research data. This research is supported by an Australian Government Research Training Program Scholarship.

†Corresponding author, Email: *baoqing.gan@uts.edu.au*; Phone: +61 2 9514 7787; Postal address: Level 7, 14-28 Ultimo Rd, (Building 8, UTS Business School), Ultimo, NSW 2007, Australia. This research constitutes a part of the corresponding author's PhD thesis. Corresponding author is grateful to the continuing support, guidance and encouragement from the supervisory panel: Vitali Alexeev *vitali.alexeev@uts.edu.au*, Ron Bird *ron.bird@uts.edu.au*, and Danny Yeung *danny.yeung@uts.edu.au*.

"Public sentiment is everything. With public sentiment, nothing can fail. Without it, nothing can succeed."

—Abraham Lincoln

1 Introduction

The financial sentiment literature has shown that macroeconomic announcements, major geopolitical events, and corporate announcements change investors' sentiment and often influence stock prices. Traditionally, investors receive this information through mainstream financial news reports, official announcements, corporate conference calls, and analysts research reports. Recent advancements in digital and telecommunication technologies facilitated social media platforms such as Twitter and StockTwits in becoming an instant channel for stock information sharing[1], disseminating greater quantities of company related information to the market at faster speeds. The importance of social media in the information dissemination process has been recognized by both regulators and market participants. For example, Bloomberg announced that it would add Twitter accounts to its financial information terminals - a "must-be" tool used by traders on Wall Street.[2] For its part, the US Securities and Exchange Commission (SEC) issued a guidance in 2008 admitting that corporate websites can serve as an effective means for disseminating information to investors, the SEC pointed out in its investigation report toward Netflix that "company communications made through social media channels could constitute selective disclosures and, therefore, require careful Regulation Fair Disclosure (Reg FD) analysis". Then in June 2015, the SEC further announced that "a start-up firm can post Twitter message about its stock or debt offering to gauge interest among potential investors" (Bartov et al., 2018), marking, for the first time, its official acceptance of social media as information dissemination channel.

Classical asset pricing models assume that investors mutually influence each other only through market price mechanisms. This assumption is less realistic since it overlooks the social interactions between investors. In reality, investors communicate and learn information through a combination of news media and social media, making social influence a critical factor of the information dissemination process and asset pricing (Hirshleifer and Teoh, 2009). Social media has been known to create attention-grabbing hot topic that may sway investors' beliefs about company's future outlook, thus forming investor sentiment that ultimately affects stock prices. For example, on 23 April 2013, a fake tweet from official Twitter account of the Associate Press announced that President Obama was injured in two explosions in the White House.[3] According to Washington Post, this hacked tweet was retweeted 4,000 times in less than five minutes with its nearly 2 million followers. Dow Jones Industrial Average (DJIA) dropped 143.5 points within 2 minutes, temporarily losing an estimated US$136 billion in value. This incident triggered critiques that the financial industry may have relied too heavily upon trading algorithms that are based on social media content.

In this paper, we explore this rapidly changing social and news media landscape that is responsible for the dissemination of information so vital to the efficient functioning of the financial markets. First, we explore the evolving relationship between social media and news media from

[1] Stafford, P. (2015), 'Traders and investors use Twitter to get ahead of market moves', *FINANCIAL TIMES*, April 29, accessed 12 August 2018, <https://www.google.com.au/amp/s/amp.ft.com/content/c464d944-ee75-11e4-98f9-00144feab7de>.

[2] Alden, W. (2013), 'Twitter arrives on the Wall Street, via Bloomberg', *The New York Times*, April 4, accessed 12 August 2018, <https://dealbook.nytimes.com/2013/04/04/twitter-arrives-on-wall-street-via-bloomberg/>.

[3] Fisher, M. (2013), 'Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism?' *The Washington Post*, 23 April, accessed 12 August 2018, <https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/?utm_term=.5e2044c627e4>.

2011 to 2017. Using the sheer volume of social/news media activity, commonly known as buzz, we documented three distinct regimes. We find that between 2011 and 2013 the news media coverage stimulates activity in social media. This is followed by transition period where news and social media activities tend to intertwine. From 2016, however, changes in quantities of social media activity seem to lead and generate news coverage volumes. We find a similar evolving pattern of lead-lag relationship between sentiment measures constructed from the tonality contained in textual data from social and news media posts.

Secondly, given that social media played a more prominent role after 2016 while news media used to be predominant before 2014, we set out to investigate the dynamic in the relationship between media activities and the stock market before and after this transition. In particular, we are interested in how news and social media sentiment affects stock returns and volatility in the periods from 2011 to 2013 and from 2016 to 2017. In dealing with inevitable endogeneity issue in the analysis of this kind, we account for the reverse influence from the stock market on social and news media. Facilitated by restricted bivariate VAR models that contain a media variable and a market variable, we find that the reaction of media sentiment to stock market shocks is more pronounced than the sensitivity of return/volatility to changes from media sentiment. This result is in line with Sprenger et al. (2014b) and Araújo et al. (2018), which find that the market features (return, trading volume and volatility) have stronger effects on media features (bullishness and posting volumes). The analysis of impulse response functions from models in the two separate periods identified above reveals that the speed of reactions for both return and sentiment have accelerated after 2016 compared to the period before 2014. Return responses to social media sentiment almost doubled after the transition period (from 0.03 to 0.07), while return responses to changes in news-based sentiment almost halved to its pre-transition level (from 0.030 to 0.016). These results corroborate our prior findings that social media is more prevalent after 2016. In contrast to return and media sentiment interactions, we find that volatility in both pre-transition and post-transition periods display higher sensitivity to social media sentiment than to news-based sentiment. Stock volatility reactions to shocks from media sentiment are more persistent than return responses. We conclude that the media sentiment does not follow market activity passively, but is actively engaging in shaping the market movements under different information environments.

Our contribution is threefold. Firstly, by separating social and traditional news media, we obtain insights into the time-varying relationship between the two information channels. As a result of the advancement in information and telecommunication technology, as well as the acceptance of the new technology by regulatory authorities, we observe propagation of social media in the later sample periods. Our results suggest that researchers in this topic should and must consider the time-varying nature of the social/news media interplay. To the best of our knowledge, there is no other research that highlights such differences, and details sentiment effects on stock market from different media sources. Secondly, accounting for the bilateral causality between media sentiment and stock market variations, we provide empirical evidence to the expanding literature on investor sentiment and noise trader risk (De Long et al., 1990). Unlike previous work, we use sentiment measures based on textual analysis that synthesizes multiple media channels' information, rather than focusing on a single platform. Lastly, our detailed statistical analysis of the Thomson Reuters MarketPsych Indices (TRMI) data adds value to the

validity of textual data in asset pricing applications by shedding light on how information from various media sources is incorporated into stock prices and volatility.

The rest of the paper proceeds as follows: Section 2 reviews previous work on investor sentiment, Section 3 describes sample data, elucidates the data pre-processing approach, and discusses research methodology, Section 4 reports results on the news and social media interplay over time, Section 5 analyses causal effects between media sentiment and stock market return/volatility. We conclude in Section 6 and propose directions for future research.

2 Literature Review

Information have played a central role in investors' choices since the advent of financial markets. Traditionally, the news media played a dominant role. For example, Tetlock (2007) shows that the content of the influential Wall Street Journal, and in particular its tone, can influence the volume of market trading and returns. Stories about a company are so important that the absence of news coverage could impact stock prices. Chan (2003) shows that companies that had no news stories experience reversals in prices in the following month. With the advancement in technology and the rise of social media, the news media is no longer the sole source of information in general and, in particular, for investors. Kwak et al. (2010) examined 106 million tweets and found that 85% of the tweets were news related. In the field of finance, studies have shown that postings on internet message boards (see Wysocki (1998), Antweiler and Frank (2004), Das and Chen (2007), and Chen et al. (2014)), and on other social media platforms such as Twitter and StockTwits (e.g., Sprenger et al. (2014a), Ranco et al. (2015)) can exert influences on stock prices and volatility. The question is whether social media have changed the way that investors consume news.

This is, literally, a billion-dollar question as social media can be subject to manipulation. For example, Lee et al. (2015) found that management use social media to mitigate the negative stock price impact of bad news about a company such as product recall. This strategy is particularly effective because it is not only the content but the tone of the message that can generate sentiment, which can influence the investor reactions to company's announcement.

Investor sentiment is the prevailing attitude of investors as to anticipated price development. It is the accumulation of variety of fundamental factors and technical indicators, such as price history, ratings and reviews, economic news reports, national and world events. According to Baker and Wurgler (2007), investor sentiment is defined as "a belief about future cash flows that is not justified by facts at hand". Broadly, investor sentiment studies can be categorised by the sentiment measure they employ: measures based on fundamental market variables, sentiment extracted from various textual sources, and sentiment scores provided by proprietary vendors such as Thomson Reuters MarketPsych and RavenPack.[4]

[4]There are categories of studies that we omit here for brevity, but nevertheless, presenting interesting directions, namely studies based on internet search behaviour, and studies relying on non-economic factors, such as weather and health conditions affecting investors' risk aversion and trading behaviour.

Investor Sentiment and Stock Market

Early research on investor sentiment and stock market movements are generally based on sentiment created from market fundamental variables (Baker and Wurgler, 2006, 2007). These sentiment proxies allowed to test behavioural finance theories such as security market under- and overreactions.[5] Empirical research on this topic makes three assumptions. First, two groups of investors play together in the market: irrational noise traders and rational arbitragers. Second: noise traders' sentiment-driven characteristics create risks to their counterparts to bet against them, which demotivate the arbitragers' trading behaviour during high sentiment periods (De Long et al., 1990). Third: there are costs to arbitrage, e.g. limit to short-sale and capital constraint (Shleifer and Vishny, 1997).

Market microstructure literature assists in tying investor sentiment to market volatility by dissecting the trading frictions, or bid-ask spread, into different components. Depending on which component is dominant, there are two mechanisms prescribing the relationship between sentiment and market volatility. First, investor sentiment negatively impacts on bid-ask spread and trading price volatility. Glosten and Milgrom (1985) proposes that adverse selection costs, as part of the bid-ask spread, are negatively correlated with sentiment-driven noise trading. In strong emotional periods, more noise trading results in narrower bid-ask spread, which concerns trading costs and risks, and price volatility. Second, investor sentiment positively influences bid-ask spread and price volatility. Order processing costs and inventory costs, taking a larger component of bid-ask spread than the adverse selection component (Huang and Stoll, 1997), are proved to be positively related to price risks and the opportunity cost of holding securities (Amihud and Mendelson, 1986). Such risk is shown to be positively linked with investor sentiment as it is harder to evaluate the misvaluations during high sentiment periods (De Long et al., 1990).

Empirical studies applying fundamental variable based sentiment index to examine stock price movement include: De Bondt and Thaler (1985), Brown and Cliff (2004), Baker and Wurgler (2006), Baker and Wurgler (2007), Barber and Odean (2007), Karlsson et al. (2009), Canbaş and Kandır (2009), Stambaugh et al. (2012), and Sayim and Rahman (2015). Findings from these studies, however, are mixed. For example, Brown and Cliff (2004) and Oliveira et al. (2013) find little or no predictability to short-term stock returns from investor sentiment, while others reveal evidence supporting the short-term price deviations as demonstrated by behavioural models.

If such short-term deviations exist, fundamental variable based sentiment indices, constructed at most monthly, may be too aggregated. More granular sentiment data at higher frequencies can be derived from other sources, providing a more detailed account of short-term fluctuations.

Investor Sentiment Based on Textual Analysis

In recent decades, advancements in textual analysis and machine learning techniques had shifted the focus of investor sentiment literature to the analysis of the relationship between stock market and information quantity, as well as sentiment conveyed within textual data (see Tetlock, 2007;

[5] Such "behaviour augmented" models usually consider various investor heuristic bias, for example, overconfidence and self-attribution bias (Daniel et al., 1998), conservatism and representativeness (Barberis et al., 1998), and confirmation bias (Rabin and Schrag, 1999)). Other behavioural models that focus on investor attention (Odean, 1999; Barber and Odean, 2007; Karlsson et al., 2009) or account for the interactions between different types of investors (Hong and Stein, 1999) have also been widely applied.

Tetlock et al., 2008; Loughran and McDonald, 2011b). Empirical research relying on scanning and scoring texts from filed documents and press releases is abundant and still expanding. There are four main information sources examined by research: **corporate filings** (e.g., Loughran and McDonald, 2011a; Jegadeesh and Wu, 2013), **professional financial news releases** (e.g., Antweiler and Frank, 2006; Engelberg, 2008; Fang and Peress, 2009; Engelberg et al., 2012; Garcia, 2013), **internet message boards** such as *Yahoo!Finance*, *RatingBull* and *SeekingAlpha* (see Wysocki, 1998; Antweiler and Frank, 2004; Das and Chen, 2007; Chen et al., 2014), and **social media platforms** such as *Twitter* and *StockTwits* (e.g., Sprenger et al., 2014b; Ranco et al., 2015), *Google* search volume (e.g., Da et al., 2011) and *Facebook's* Gross National Happiness index (Siganos et al., 2014).[6]

Most of the empirical work focuses on either the volume (e.g., coverage) or the sentiment (positive vs negative emotions or tonality) conveyed in textual data, research that considers both is rarely observed. In fact, as pointed out by Liu and McConnell (2013), both the level of media attention and the tones within press articles are significantly associated with the various types of corporate events, which ultimately impact stock prices and volatility. We adhere to this view and conduct our analysis accounting for both the level of coverage and the sentiment tonality expressed by media outlets.

Due to the limited computational power at early stages of textual analysis and the requirement of manually-handled "training" process for algorithms such as Naive Bayesian Classification, sample sizes in some of the earlier works are relatively small. One could only focus on either a small group of representative companies, or constrain the sampling period to a short time frame, but not both.[7] This small sample problem is better dealt with in Leung and Ton (2015) and Renault (2017). Covering more than 2,000 public firms in Australia from 2003 to 2008, Leung and Ton (2015) examines over 2.5 million stock related messages posted on *HotCopper* forum, and finds that small, high growth, and hard-to-valuation stocks tend to be easily affected by internet message board. Renault (2017) abstracts textual sentiment from 750,000 StockTwits at intra-day level between September 2014 and April 2015 and finds that sentiment changes in the first-half trading hour manifest market return predictability to the last half-hour.

Investor Sentiment Based on MarketPsych Indices

To break the confinements of data availability from small number of assets, short observation period, and single type of media source, several studies reap the reward of unique data set from professional financial data vendors such as Thomson Reuters and Dow Jones. This type of data takes advantage of combining more comprehensive content for certain categories of information (news or social media), rather than focusing on a standalone platform. For instance, using sentiment indicators from Thomson Reuters News Scope (TRNS) and texts data from Thomson Reuters News Archive (TRNA), Heston and Sinha (2017) validates the effectiveness of textual sentiment data to predict stock returns. They provide evidence that daily textual sentiment only predict return at short-term (one or two days) horizon, whereas weekly sentiment indices

[6] Our review of empirical research that utilize various textual data sources in this field is far from exhaustive. For comprehensive survey, refer to Kearney and Liu (2014) and Brzeszczyński et al. (2015).

[7] For example, Ranco et al. (2015) uses Twitter API to analyse 30 Dow Jones companies involving 151 events and covering the period from June 2013 to September 2014, while Das and Chen (2007) examines 24 high-tech companies in the two-months period from July to August 2011.

contains predictability up to a quarter.

Different from News Analytic data, Thomson Reuters MarketPsych Indices (TRMI), the dataset employed in this paper, contains synthesized quantities and emotional measures from a wide range of traditional news channels as well as social media platforms.[8] We contrast sentiment captured by TRMI from social and news media to the Baker & Wurgler index (BW) commonly used in investor sentiment analysis.[9] The correlations between social and news media TRMIs and the BW index are 0.54 and 0.44 respectively, demonstrating a degree of commonality between TRMI sentiment indicators and the *BW* index.[10] Yet, the magnitudes of correlation coefficients are indicative of divergence of these two measures, suggesting that the TRMI sentiment indices capture different investor sentiment from BW. On one hand, strong positive correlation provides merit for using TRMI as it captures commonality in general trend of these two indicators. On the other hand, TRMI provides sentiment scores at a much higher frequencies allowing us to study the dynamics in temporal displacement within sentiment scores (news vs social) and between sentiment and market variables (sentiment vs returns and/or volatility).

Recent studies have already shown the effectiveness and validity of this dataset in measuring media-related investor sentiment. For example, Michaelides et al. (2015) (see Table 5 therein) matches the manually collected sovereign downgrade news events with TRMI metrics, and confirms the consistency and validity of TRMI variables. A further research conducted by Michaelides et al. (2018) uses TRMI and manually constructed FX currency related news to control for media based public information, confirming consistency between these two groups of measures. Investigating the market dynamics between TRMI sentiment index and Brazil stock index (IBovespa), Araújo et al. (2018) finds strong reverse causation from market movements to media sentiment.

Our paper is complimentary to Sun et al. (2016), Nooijen and Broda (2016), and Jiao and Walther (2016) in that we focus on the aggregate US equity market. Concentrating on intraday (half-hour) data from TRMI, Sun et al. (2016) explores the within day return predictability for the Index. They substantiate that changes of TRMI sentiment in the first half trading hour are helpful to forecast the last two trading hours' stock index returns, which is different from within day momentum effect. They point out that this predictability enables to create economic value when evaluated with market-timing strategy. Examining the MSCI US Equity Sector Indices from TRMI, Nooijen and Broda (2016) finds higher predictability for stock volatility than for return. They highlight the significance of distinguishing different market environments, for example, calm or volatile periods. Contrasting social media with news using TRMI media quantity measures, Jiao and Walther (2016) develops a generalised asset pricing model that accommodates various behavioural biases. They use this model to examine social and news media effects on volatility and volume of 2,613 US stocks from 2009 to 2014. They document evidence that higher social media sentiment leads to higher volatility and trading volume in the next months. In contrast, improvements in news sentiment result in decreased volatility and volume in the coming month.

[8] While description of the sub-sample employed in this paper is presented in Section 3, the detailed summary of the full dataset is provided in the supplementary online appendix.

[9] We are grateful to Jeffrey Wurgler for making their monthly investor sentiment data publicly available on his website at NYU Stern. Assessed on 8 February 2019, <http://people.stern.nyu.edu/jwurgler/>.

[10] See supplementary online appendix for details.

This paper contributes to the literature in several ways. Firstly, similar to Jiao and Walther (2016), we discriminate two different types of media, social vs news, and examine the dynamics in the lead-lag relationships between these two channels from both the activeness (*Buzz*) and the emotions (*Sentiment*) conveyed in data from these two channels. But, in contrast to Jiao and Walther (2016), we address the important question: had the media landscape changed from 2011 to 2017, and how social and news media had interacted with each other over this period. Secondly, as pointed out by Baker and Wurgler (2007) and Nooijen and Broda (2016), we emphasise the importance of time-varying relationship between investor sentiment and the market. That is, we analyse the mutual causality between media sentiment and stock market variables (return and volatility) under different market information environments: (i) period of conventional news media dominance, (ii) transitory period with no clear lead effect of one information channel over the other, and (iii) period of increasing dominance of social media. Extending the strand of literature that uses MarketPsych Indices investor sentiment, our exploration and results reveal new facts about the role of information in asset pricing in the social media era.

3 Data and Methodology

Our dataset is comprised of two sources: sentiment data and stock market data. Our sentiment data is based on Thomson Reuters MarketPsych Indices (TRMI) textual analysis scores for the company group. Our stock market data are obtained from Datastream and Wharton Research Data Services. Details on each dataset and data pre-processing methods are provided below.

3.1 Sentiment Data

In contrast to the definition in Baker and Wurgler (2006), we refer to investor or market sentiment as the overall attitude of investors toward a single security or financial market. It is the tone of an asset or a market, its crowd psychology. Thomson Reuters MarketPsych Indices (TRMI) incorporates analysis of news and social media in real-time by translating the quantity and emotions of financial economic news and internet messages into manageable information flows.[11] TRMI provides three content categories: **news**, **social** and **combined**, based on English language articles and posts dating back to 1998. TRMI covers more than 2,000 news sources, including leading professional financial news presses such as *The Wall Street Journal, The Financial Times, and The New York Times*, as well as other less influential news content synthesised by Thomson Reuters News Feed Direct, Factiva News, *Yahoo!* and Google News. TRMI also claw and scrape the top 30% of over 2 million blogs, stock message boards and social media sites minute-by-minute, including StockTwits, *Yahoo!Finance*, and *SeekingAlpha*. Term weighting and scoring approach of TRMI is based on the Loughran and McDonald (2011b) dictionary scheme, which is proved to be more suitable to financial contexts rather than the psycho-social dictionary scheme of the Harvard General Inquirer (GI) used in Tetlock (2007). These data allow us to study and contrast the difference in sentiment effects from social and news media.

[11] The data are provided by Thomson Reuters Financial and Risk Team as part of TRMI product. TRMI covers a plethora of securities and markets, including: more than 12,000 companies, 36 commodities and energy subjects, 187 countries, 62 sovereign markets, 45 currencies, and, since 2009, more than 150 cryptocurrencies. For more details, see *Thomson Reuters MarketPsych Indices 2.2 User Guide, 23 March 2016, Document Version 1.0.*

TRMI offers three types of sentiment indicators for a specific company or company group: 1) **Emotional** indicators including *Sentiment*, *Anger*, and *Fear*; 2) **Fundamental** perceptions such as *Long vs Short*, *Earnings Forecast*, and *Interest Rate Forecast*; and 3) **Buzz** metric, a measure indicative of how much activity market-moving topics, such as *Litigation*, *Mergers*, and *Volatility* are being generated and discussed. After the social media posts or news articles are published in the TRMI content sources, a linguistic software abstracts the new content feed, parses and scores the content and attributes the score to global indices, companies, bonds, countries, commodities, currencies, and cryptocurrencies. In fact, TRMI offers a total of 35 emotional scores. We decide to focus on *Sentiment* after performing Principal Component Analysis (PCA) and checking variance decomposition of the first two principal components.[12] However, *Buzz* metric is conceptually different from the emotional and fundamental scores. It measures the volume of information flow and, therefore, is not incorporated in the PCA analysis with other scores. Yet, *Buzz* metric is crucial in our analysis of social vs traditional news media dominance throughout the sample period.

Several studies have verified the validity of the textual sentiment measures provided by TRMI e.g., Michaelides et al. (2015), Sun et al. (2016), Nooijen and Broda (2016), and Michaelides et al. (2018). In our analysis we employ daily observations from 2011 to 2017 for the *MPTRXUS500* company group index that aggregates sentiment and tone of the largest 500 companies in the US, and aims at capturing the index sentiment. The data are updated each day at 3:30pm US Eastern time, including weekends and other non-trading days.[13] According to Heston and Sinha (2017), daily textual sentiment possesses short-term return predictability. Table 1 presents descriptive statistics for the sentiment indices and the media activity measure, *Buzz*, based on social media and news respectively. Sentiment scores are buzz-weighted, averaging any positive references net of negative references in the last 24 hours. Upon examination of the descriptive statistics, we observe the following facts: first, *Buzz*, a sheer media coverage volume metric for both social and news media, has a much larger absolute value than sentiment (average *Buzz* value of 116,484.46 for social media and 202,401.31 for news, while sentiment mean values are close to zero). Social media *Buzz* is highly positively skewed with the third moment equals to 1.37, and contains several large outliers.The kurtosis of 6.32 indicates a leptokurtic distribution. In contrast, news media buzz is more symmetric and contains less outliers than social media, with skewness equal to -0.01 and kurtosis 3.91 - slightly higher than 3. Lastly, all of the TRMI indices are significantly autocorrelated with potential long memories.[14]

3.2 Stock Market Data

The sample period for the stock market data is consistent with the availability of our TRMI data and sampled daily from January 1, 2011 November 30, 2017. Fortunately, this period avoids the turmoil of the global financial crisis (GFC) episodes from 2008 to 2010 and escapes potential influence of change in data sources last reported by TRMI in 2009. At the same time, this sample period covers a phase of rapid development of social media, allowing us to compare and contrast

[12] Results of our PCA analysis are detailed in the supplementary online appendix.
[13] Further details on the TRMI data can be found in the MarketPsych white paper by Peterson (2013).
[14] In the unreported tables, we conduct Durbin-Watson (DW) test and Ljung-Box test with up to 5 lags (LB-5). Evidence of autocorrelation with potential long memories for all available social and news emotional indices are available upon request.

social and news based sentiment directly. Following Antweiler and Frank (2004), and Sprenger et al. (2014b), we employ stock return and volatility as our main stock market variables, with descriptive statistics summarised in Table 1:[15]

Table 1: Descriptive Statistics for the company group over the period 2011/01/01-2017/11/30. *Sentiment*, obtained from TRMI, is bounded on [-1,1]. Negative and positive values denote negative and positive sentiment, zero denotes neutral score. *Buzz*, representing the volume of information flow, differs from *Sentiment* index, and is only bounded from below at 0. *Sentiment* and *Buzz* indices are obtained from TRMI under asset group code MPTRXUS500 which aggregates information on the top 500 US-based companies and resembles Index. *Returns* are calculated as $r_t = log(P_t/P_{t-1})$, where P_t is the daily close price for the index obtained from Datastream. Reported return figures are annualized by multiplying the daily return values by 252. *VIX* data is acquired from WRDS CBOE volatility index futures closing prices. The unreported Durbin-Watson test and Ljung-Box 5 lags test for all indices show presence of autocorrelation for all series.

	Mean	Std	Max	Min	Skew	Kurt	25th	Median	75th	IQR
Social media:										
Sentiment	-0.020	0.030	0.082	-0.127	-0.32	2.80	-0.040	-0.016	0.001	0.042
Buzz	116,484	35,769	311,543	14,179	1.37	6.32	94,587	110,860	130,317	35,730
News media:										
Sentiment	-0.017	0.037	0.126	-0.173	-0.29	3.22	-0.042	-0.015	0.009	0.051
Buzz	202,401	47,847	387,635	1,468	-0.01	3.91	172,081	202,994	231,451	59,369
Market:										
Return	0.09	1.99	10.42	-15.52	-0.54	8.78	-0.68	0.06	1.07	1.75
VIX	16.34	5.58	48	9.14	2.07	8.34	12.85	14.89	17.96	5.11

We believe that the implied volatility of stock index futures (VIX) is more suitable to our analysis than the traditional realised volatility measures since investor sentiment is tied to a forward looking perspective, as defined by Baker and Wurgler (2007). On the contrary, realised volatility such as standard deviation or squared terms of prior period returns, takes a backward looking view, and thus is less relevant to our investigation. This is in line with Han and Park (2013) who compares realised volatility and VIX and proves the appropriateness of VIX for out-of-sample and forward-looking research.

Our econometric frameworks requires that variables are covariance stationary, with their first two moments finite and time-invariant. Our results from unit root tests indicate that all variables are covariance stationary.[16]

3.3 Data Aggregation Process

In order to familiarise the reader with the properties of our two main TRMI indices, *Buzz* and *Sentiment*, we plot the raw series, autocorrelation functions (ACF) and partial autocorrelation functions (PACF) up to 40 lags in Appendix Figures A.1 and A.3 (pages 31 and 32). We observe large outliers and strong weekly seasonality in *Buzz* series for both social and news media. Winsorizing *Buzz* metrics at the 99 percentile (right tail only) mitigates the effects of extreme outliers.[17]

To deal with weekly effects in *Buzz* and *Sentiment* series, we regress *Sentiment* and win-

[15] A full list of all data sources and acronyms is available in Table A.1 in the appendix.

[16] Augmented Dickey-Fuller and Phillips-Perron unit test results for models with a (i) constant, (ii) drift, and (iii) drift and time trend are presented in Section A.2 of the appendix.

[17] We perform asymmetric winsorizing since *Buzz*, describing media activity quantities, is bounded on $[0, \infty)$.

sorized *Buzz* on day-of-the-week dummy variables, retaining fitted residuals as our seasonally adjusted data. Figure A.2 in the appendix plots the winsorized and seasonality adjusted *Buzz* series. Lastly, we align seasonality adjusted TRMI indices with market variables for trading days only. The values for sentiment indices during non-trading days are averaged with the sentiment index value on the first trading day immediately after a weekend or public holiday. For example, sentiment indices on Monday represent average values based on Saturday, Sunday and Monday sentiment scores. Figure A.4 in the appendix depicts the seasonality adjusted and non-trading day merged *Sentiment* series. After combining with stock market data, our sample size reduces from 2,526 observations to 1,803 for each time-series. A comparison of Figures A.2 and A.4 shows that we have successfully removed the weekly seasonality from both the buzz and sentiment series. This concludes our data pre-processing, with both series, *Buzz* and *Sentiment*, exhibiting stationary, strong autocorrelation and long memory, allowing us to pinpoint the best econometric framework for this type of series.

3.4 Econometric Framework

To capture interdependence between news and social media while avoiding explicit exogeneity assumptions, we adopt the vector autoregressive (VAR) framework.[18] VAR provides a simple framework systematically capturing rich dynamics in multiple time-series. We rely on a rolling-window VAR approach to investigate our main research questions, respectively: (1) How social and news media interact with each other over time? (2) What are the dynamic relationships between media activities and stock market activities?

To identify a group of simultaneous equation models, one has to make assumptions about endogeneity of the variables considered: which variables are deemed endogenous while others are purely exogenous? These decisions are often criticized as being too subjective (Gujarati, 2009). VAR overcome this shortcoming since it does not assign any prior distinction between endogenous and exogenous variables, i.e. all variables in VAR are endogenous. Thus, to investigate how social and news media activeness (*Buzz*) and emotions (*Sentiment*) intertwine with each other over time, and further to probe how media sentiment and stock market associate with each other, we adopt a general VAR framework setup shown as follow:[19]

General Setup: Let $\mathbf{x}_t$ be a multivariate time series, a VAR process of order 1, or VAR(1) for short, follows the model:

$$\mathbf{x}_t = \phi_0 + \mathbf{\Phi} \cdot \mathbf{x}_{t-1} + \epsilon_t$$

where ϕ_0 is a k-dimensional vector, $\mathbf{\Phi}$ is a $k \times k$ matrix, and $\{\epsilon_t\}$ is a sequence of serially uncorrelated random vectors with mean zero and covariance matrix Ω.[20] For instance, $\mathbf{x}_t$ could consist of any number of the following variables:

- market data (e.g., *return*, *volume*, and/or *volatility*);
- TRMI social indices (e.g., *buzz*, *sentiment* and/or *fear*);
- TRMI news indices (e.g., *buzz*, *sentiment*, *gloom*, etc.);

[18] Sims (1980) advocated VAR models as providing a theory-free method to estimate linear interdependence among time-series and to avoid the "incredible identification restrictions".
[19] A full list of variables, the notations and definitions of them used in this study is available in Table A.1.
[20] $\{\epsilon_t\}$ is also called impulse, or innovations (Tsay, 2005).

$\mathbf{x}_t$ can be generalized to VAR(p), where p is the number of lags considered. To choose the appropriate lag length, p, we use the Akaike Information Criterion (AIC) and Schwartz's Bayesian Information Criterion (BIC).[21] BIC generally penalizes free parameters more strongly than AIC, allowing for more parsimonious models.[22]

4 News vs Social Media: Dominating Causality Pattern

We investigate the question: how financial news media landscape changed in the past decade by examining the dynamic relations between $Buzz_S$ and $Buzz_N$ from estimating a VAR(1) model using S&P500 TRMI company group data. $Buzz$ metric is conceptually different from the emotional and fundamental scores.[23] It measures the volume of information flow and, in fact, is used in calculating the emotional and fundamental scores. We choose it because it is the most representative stock index in the US market, comprising of the most liquid large-cap companies representing approximately 80% of the US equity market capitalization. By restricting the analysis to the S&P500 group, we ensure that the companies in our aggregate sample are sufficiently large to receive regular media coverage. To help with the interpretation of the results, we rewrite the general VAR model in scalar form, where we set $k = 2, \mathbf{x}_t = (Buzz_S, Buzz_N)'$:

$$\begin{aligned} Buzz_{S,t} &= \phi_{S,0} + \Phi_{1,1} Buzz_{S,t-1} + \Phi_{1,2} Buzz_{N,t-1} + \epsilon_{1,t}, \\ Buzz_{N,t} &= \phi_{N,0} + \Phi_{2,1} Buzz_{S,t-1} + \Phi_{2,2} Buzz_{N,t-1} + \epsilon_{2,t}. \end{aligned} \tag{1}$$

Here, $\Phi_{1,2}$ denotes the linear dependence of $Buzz_{S,t}$ on $Buzz_{N,t-1}$ with lagged dependent variable $Buzz_{S,t-1}$ also as a regressor, so $\Phi_{1,2}$ captures the conditional effect of $Buzz_{N,t-1}$ to $Buzz_{S,t}$ given $Buzz_{S,t-1}$. Analogous interpretation for $\Phi_{2,1}$ applies. Gujarati (2009) distinguishes four cases for such VAR system:

1. Unidirectional causality from $Buzz_N$ to $Buzz_S$ if $\Phi_{1,2}$ is significantly different from zero while $\Phi_{2,1}$ is **NOT** significantly different from zero;
2. Inverse unidirectional causality from $Buzz_S$ to $Buzz_N$ if $\Phi_{2,1}$ is significantly different from zero while $\Phi_{1,2}$ is **NOT** significantly different from zero;
3. Feedback, or bilateral causality, when **both** $\Phi_{1,2}$ and $\Phi_{2,1}$ are significantly different from zero;
4. Independence, when **neither** $\Phi_{1,2}$ nor $\Phi_{2,1}$ are significantly different from zero.

Our interest lies in the off-diagonal regression coefficients because the level and significance of VAR off-diagonal coefficients characterize causal relationships, while diagonal elements only show autocorrelation effects.

To perform a rolling-window analysis, we use the past 365 days (i.e. the prior one-year period) as an estimation window. We obtain off-diagonal elements of slope coefficients (Φ_{12} and Φ_{21}) and test their significance. We repeat this analysis on each day for the reminder of the sample to capture the dynamics and evolution of the causal relationship over time. Figure 1 presents the results of this procedure. Each vertical pair of observations represents the off-diagonal slope coefficients of a VAR(1) model. Statistically significant results are emphasised

[21] For notation and definition details, refer to Table A.1 in the appendix.
[22] We conduct formal hypothesis test using the likelihood ratio statistic. Our results are reported in Section B.2.3.
[23] See supplementary online appendix for details on these definitions in Section B.3.

with bold points.[24] Following DeMiguel et al. (2014), we define "dominating" or "leading" series as follow: in an off-diagonal coefficients plot of a two-variable rolling-horizon VAR system, if one coefficient is significant, the other coefficient is insignificant, then the significant series "leads" or "dominates" the insignificant series. If both coefficients are significant, then the higher magnitude coefficient "leads" or "dominates" the lower magnitudes series.

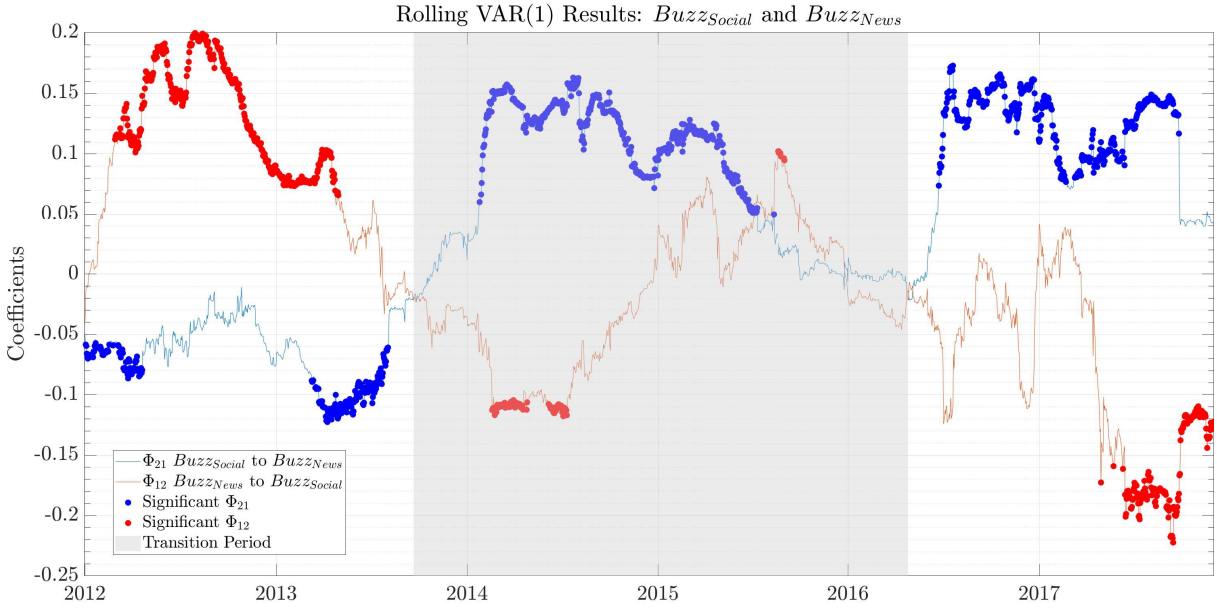


Figure 1: Rolling Window VAR(1) Off-Diagonal Elements - daily *Buzz*. This plot depicts the inter-relationships between $Buzz_S$ and $Buzz_N$ series from 2011/01/01 to 2017/11/30. Sample contains 2,526 observations for each series, with the first 365 observations used as pre-estimation window. The red line represents the leading effect from news media to social media, Φ_{12} in equation system (1), and the blue line indicates the leading effect from social media to news, Φ_{21} in equation system (1). Coefficients that are significant at the 90% level are shown with bold dots. The shaded area indicates a transition period. We rely on the crossings of the two lines: estimated effect of previous $Buzz_S$ on current $Buzz_N$ (blue line) and estimated effect of previous $Buzz_N$ on current $Buzz_S$ (red line). In the beginning of our sample, as we roll the estimation window, the two lines converge. Towards the end of our sample, the two lines begin to diverge. We used the first and the last crossing points as the dates for the beginning and the end of the transition period, respectively.

From Figure 1, we observe that the blue and red coefficients crossed in October 2013. Prior to this "transition" point, the magnitude of red line (Φ_{12}) is above blue line (Φ_{21}), with more numbers of Φ_{12} coefficients being significant than the Φ_{21} coefficients. For example, in Table 2 Panel A left side, we report one of the VAR(1) results based on equation (1) in the pre-transition period. ϕ_{12}, the impact from $Buzz_N$ to $Buzz_S$, is 0.1927 and significant at 1% level. By contrast, ϕ_{21}, the impact from $Buzz_S$ to $Buzz_N$, is -0.0329 and not statistically significant. This phenomenon reveals the fact that news media activity dominates social media activities before October 2013. After this "flip-point", we observe that the values of blue coefficients exceed the red coefficients. From 2014 to 2016, there are periods that both blue and red coefficients are significant, indicating news and social media mutually Granger cause each other. We interpret this

[24] Based on our analysis, a VAR model with 7 lags is optimal according to BIC criterion. Detailed AIC and BIC results are available in our supplementary appendix (Table B.2 Panel A on page 35). However, we report VAR(1) as it is a parsimonious form of VAR(7) based on the model specification test shown in Table B.3 on page 36 of the appendix. According to Table B.3, most of the intermediate lags' coefficients in VAR(7) model are insignificant, and only the coefficients of the seventh-lag and the coefficients of the first lag are significant, suggesting that the optimal lags determined by the information criteria might be due to the remaining weekly seasonality, which could not be modelled. Similar rolling window VAR(1) approach was used in DeMiguel et al. (2014) in investigating the cross-correlations between size portfolios over time. The results of our VAR(7) model are available upon request.

period as a transition period (the grey shaded period). We find that the "flip-point" date identified from our data coincidences with the SEC's permission to new format media announcements as mentioned in Section 1. Lastly, we find that after mid-2016, Φ_{21} (the blue line, social to news) trends further upward, remaining significant, while Φ_{12} (the red line, news to social) fluctuates and tend to trend downward, indicating a prominent influence of social media on conventional news. Meanwhile, as shown in the right side of Panel A Table 2, ϕ_{21}, the coefficient from $Buzz_S$ to $Buzz_N$, equals to 0.1101 and is significant at 1% level, while a lower level ϕ_{12}, the coefficient from $Buzz_N$ to $Buzz_S$, is not statistically significant. This result confirms the dominant effect of social media over news after January 2016. Overall, our results shows that there has been a change in the information landscape and market conditions with the distinct propagation of social media is playing a predominant role in the flow of information.

Table 2: Before vs After Transition Period VAR Slope Coefficients: Social vs News. Panels A and B report the estimated VAR(1) slope coefficients for system equations (1) and (2) respectively. p-values below 0.1, 0.05, and 0.01 are denoted as *, **, and *** respectively. In Panel A, ϕ_{12} represents the effects from news media volume to social media activeness, while ϕ_{21} shows the impacts from social media activity to news article volume. ϕ_{11} and ϕ_{22} in Panel A are the autocorrelations for $Buzz_S$ and $Buzz_N$ respectively. In Panel B, ϕ_{12} and ϕ_{21} coefficients represent the effects from net sentiment on news media to social media based sentiment, while ϕ_{21} shows the impacts from social media sentiment to news-based sentiment. ϕ_{11} and ϕ_{22} in Panel B are the autocorrelations for $sent_S$ and $Sent_N$ respectively.

Panel A: $Buzz_S$ vs $Buzz_N$

	Pre-transition Period					Post-transition Period			
	Coef.	*s.e.*	*t*-stat	*p*-value		Coef.	*s.e.*	*t*-stat	*p*-value
ϕ_{11}	0.8719	0.0418	20.86***	0.00***	ϕ_{11}	0.5199	0.0684	7.60***	0.00***
ϕ_{12}	0.1927	0.0388	4.96***	0.00***	ϕ_{12}	0.0416	0.0998	0.42	0.68
ϕ_{21}	-0.0329	0.0547	-0.60	0.55	ϕ_{21}	0.1101	0.0435	2.53***	0.01***
ϕ_{22}	0.5577	0.0508	10.97***	0.00***	ϕ_{22}	0.7021	0.0634	11.07***	0.00***

Panel B: $Sent_S$ vs $Sent_N$

	Pre-transition Period					Post-transition Period			
	Coef.	*s.e.*	*t*-stat	*p*-value		Coef.	*s.e.*	*t*-stat	*p*-value
ϕ_{11}	0.6421	0.0465	13.81***	0.00***	ϕ_{11}	0.6807	0.0435	15.65***	0.00***
ϕ_{12}	0.2325	0.0601	3.87***	0.00***	ϕ_{12}	0.0166	0.0481	0.34	0.73
ϕ_{21}	-0.0089	0.0390	-0.23	0.82	ϕ_{21}	-0.1589	0.0470	-3.38***	0.00***
ϕ_{22}	0.4503	0.0504	8.94***	0.00***	ϕ_{22}	0.3907	0.0520	7.51***	0.00***

Next, we examine how the emotions expressed in news and social media intertwine with each other across time. Following the same methodology, we represent $k = 2, \mathbf{x}_t = (Sent_S, Sent_N)'$ in the General Setup of VAR(1)[25] In Figure 2, we observe a sharp difference in the magnitudes of VAR coefficients (between Φ_{12} and Φ_{21}) prior to the shaded transition period. Specifically, the one-day lead effect from news sentiment to social (red, Φ_{12}) is significantly higher than the effect from social sentiment to news (blue, Φ_{21}). For example, in the left side of Panel B in Table 2, one of the VAR regression results in the "Pre-transition Period" shows that the coefficient of news to social sentiment effect (ϕ_{12}) is 0.2325 with t-statistics and p-value significant at 1% level. In contrast, the coefficient of social to news sentiment effect (ϕ_{21}) is -0.0089, a much lower level compared with ϕ_{12}, 0.2325, with insignificant p-value (0.82). Continuing our investigation

[25]Table B.3 Panel B in the Appendix provides evidence substantiating that VAR(1) is a parsimonious model of VAR(7) by listing coefficient estimates for intermediate lags and their significance levels, and rewrite the model as equation system (2):

$$\begin{aligned} Sent_{S,t} &= \phi_{S,0} + \Phi_{1,1} Sent_{S,t-1} + \Phi_{1,2} Sent_{N,t-1} + \epsilon_{1,t} \\ Sent_{N,t} &= \phi_{N,0} + \Phi_{2,1} Sent_{S,t-1} + \Phi_{2,2} Sent_{N,t-1} + \epsilon_{2,t} \end{aligned} \tag{2}$$

The rolling-window results from equation system (2) are plotted in Figure 2.

of Figure 2, we find that, in spite of some fluctuations in the transition period when news and social mutually influence each other, the impact of social media sentiment effect dominates in the final part of our sample period, which is similar to the buzz analysis pattern. We also observe that most of the red (Φ_{12}) coefficients are not significant in this post-transition episodes, while more blue (Φ_{21}) coefficients are significant and at higher magnitudes. For instance, the right side of Panel B in Table 2 indicates that one of the "Post-transition Period" VAR has social to news effect (ϕ_{21}) of -0.1589, which is significant at 1% level. But influences from news to social media sentiment (ϕ_{12}) become insignificant (p-value of 0.73) at a lower level of 0.0166. This result is consistent with the pattern we identified in Figure 1. In both figures, news media impacts are leading social media effects before the transition period, however, after the transition period, this pattern is reversed.

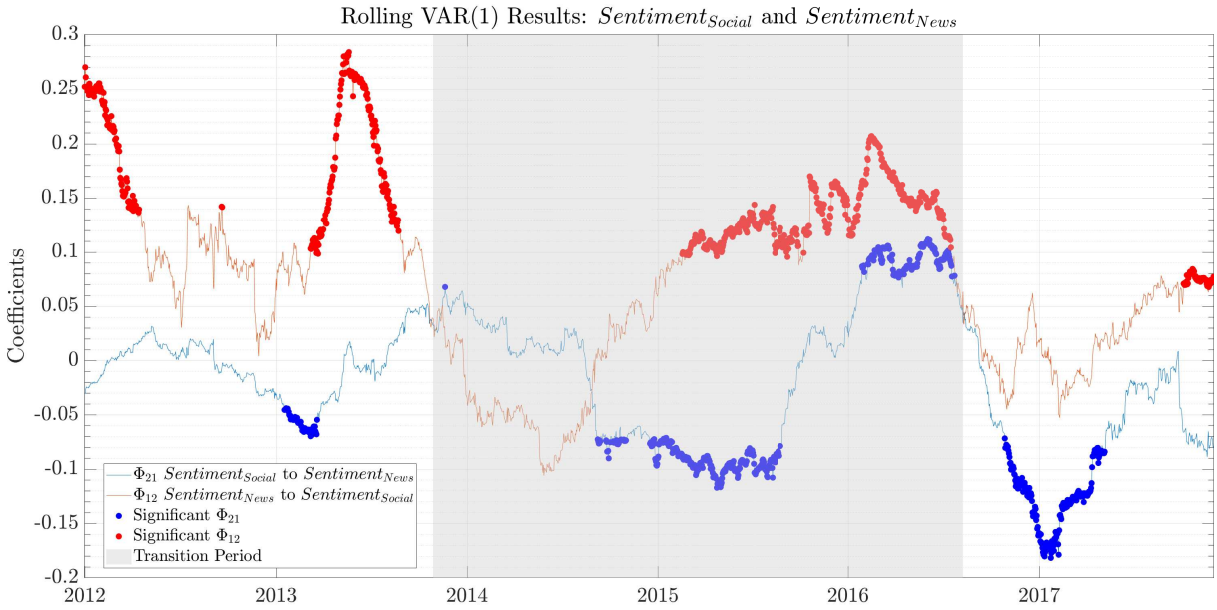


Figure 2: Rolling Window VAR(1) Off-Diagonal Elements - daily *Sentiment*. This plot depicts the inter-relationships between $Sent_S$ and $Sent_N$ series from 2011/01/01 to 2017/11/30. Sample contains 2,526 observations for each series, with the first 365 observations used as pre-estimation window. The shaded area indicates a transition period. The red line represents the leading effect from news media to social media, Φ_{12} in equation system (2), and the blue line indicates the leading effect from social to news, Φ_{21} in equation system (2). Coefficients that are significant at the 90% level are shown with bold dots.

5 Media vs Market: Sub-sampling Period Comparison

Now that we have established that there is a transition period, we turn our attention to the question of how sentiment impacts on the stock market during the two periods: the pre-2014 and post-2016 sessions. Accordingly, we merge and synchronise the seasonality adjusted social and news *Sentiment* series with stock variables by averaging *Sentiment* values on non-trading days. Next, to deal with the scale difference problem, we standardise all series to have zero mean and unit standard deviation prior to estimation. As identified in the previous section, we separate our sample period into three sub-periods: the pre-transition period (from Jan 2011 to Dec 2013), the transition period (from Jan 2014 to Dec 2015), and the post-transition period (from Jan 2016 to Nov 2017).

5.1 Sentiment vs Return

To examine the relationship between returns and sentiment, we estimate the following two systems by replacing $k = 2$, $x = (Sent_S, r)'$ and $x = (Sent_N, r)'$ respectively in the General Setup of VAR(1):

$$\begin{aligned} Sent_{S,t} &= \phi_{S,0} + \Phi_{1,1} Sent_{S,t-1} + \Phi_{1,2} r_{t-1} + \epsilon_{1,t} \\ r_t &= \phi_{N,0} + \Phi_{2,1} Sent_{S,t-1} + \Phi_{2,2} r_{t-1} + \epsilon_{2,t} \end{aligned} \tag{3}$$

$$\begin{aligned} Sent_{N,t} &= \phi_{S,0} + \Phi_{1,1} Sent_{N,t-1} + \Phi_{1,2} r_{t-1} + \epsilon_{1,t} \\ r_t &= \phi_{N,0} + \Phi_{2,1} Sent_{N,t-1} + \Phi_{2,2} r_{t-1} + \epsilon_{2,t} \end{aligned} \tag{4}$$

This VAR setup allows us to account for the mutual impacts between return and media sentiment. We focus on the pre-2014 and after-2016 episodes, omitting the transition period because the dominating pattern during the transition period is less obvious.[26].

Panels A and B in Table 3 summarise the results for VAR systems in (3) and (4) respectively over pre- and post-transition periods. The coefficients estimated are the initial sensitivities of the dependent variable to lagged independent variables. For example, ϕ_{12} from both pre- and post-transition periods in Panels A and B are positive and significant at 5% level: 0.0995 in the pre-transition period, and 0.1929 in the post-transition period for the social media sentiment VAR system in Panel A; 0.1060 in the pre-transition period and 0.2171 in the post-transition period for the news sentiment regression in Panel B. These results indicate that returns have positive and significant impacts on both social and news sentiment. In contrast, initial sensitivities of returns to sentiment, the ϕ_{21} coefficients in Panels A and B, are insignificant for all four estimators. This result is consistent with the extant literature. For example, Sprenger et al. (2014a) also finds that the feedback effect from stock market to social media variables prevails. To get a better understanding of these results, and to contrast news and social media effects, we generate Impulse Response Functions (IRFs) for the leading 20 working days (equivalent to approximately one month) in Figure 3.

Plots on the left of Figure 3 represent IRFs that capture return responses to social or news media sentiment shocks. Panels (a) and (c) represent responses of return to **social** meida sentiment shocks in the pre- and post-transition period respectively, whereas Panels (e) and (g) are return responses to **news** sentiment shocks in the two sub-sampling periods respectively. All four left-hand side IRFs show that the initial impacts on return from sentiment (both social and news) are positive, and reverting back to zero gradually with deviations at different speeds. This finding is consistent with the overreaction hypothesis, which proposes that sudden surges in investor sentiment lead to temporarily spikes in stock prices that will retreat shortly.

A comparison of Panels (a) and (c) of Figure 3 reveals two interesting findings. First, the influence of social media sentiment on return increased after the transition period. In particular, the magnitude of IRFs expands from 0.03 before 2014 to 0.07 after 2016 - the sensitivity almost

[26] As is shown in Table B.2 in the Appendix, VAR(5) is optimal for these two systems according to BIC. However, we report VAR(1) results in Table 3 due to parsimony of VAR(1) model combined with the fact that intermediate lags, that is lags 2, 3, and 4, are insignificant. The lag 5 (trading days only data) corresponds to remaining weekly seasonality, which could not be modelled. This is consistent with our analysis in Section 4, where we analysed sentiment indices and observed significance at lag 7 (calendar day weekly seasonality)

Table 3: Before vs After Transition Period VAR Slope Coefficients: Sentiment Vs Market. Panels A to Panel D report the estimated VAR(1) slope coefficients for equation systems (3) to (6) respectively. p-values below 0.1, 0.05, and 0.01 are denoted as *, **, and *** respectively. In Panel A, ϕ_{12} represents the effects from return shocks to social media sentiment, while ϕ_{21} shows the impacts from social media sentiment to return. ϕ_{12} and ϕ_{21} coefficients in Panel B represent the same lead-lag relations as shown in Panel A, but for news-based sentiment. ϕ_{11} and ϕ_{22} are the autocorrelation for *Sentiment* and *Return* in Panels A and B. Likewise, in Panel C, ϕ_{12} represents the effects from volatility (VIX) to social media sentiment, while ϕ_{21} shows the impacts from social media sentiment to stock volatility. ϕ_{12} and ϕ_{21} coefficients in Panel D represent the same lead-lag relations as shown in Panel C, but for news-based sentiment. *Sentiment* is measured as the squared term of the seasonality adjusted and non-trading day averaged *Sentiment* series in Panels C and D, and ϕ_{11} and ϕ_{22} are the autocorrelation for $Sent^2$ and VIX in these two panels.

Panel A: $Sent_S$ **vs** *Return*

	Pre-transition Period					Post-transition Period			
	Coef.	*s.e.*	t-stat	p-value		Coef.	*s.e.*	t-stat	p-value
ϕ_{11}	0.3957	0.0581	6.81***	0.00***	ϕ_{11}	0.6041	0.0503	12.02***	0.00***
ϕ_{12}	0.0995	0.0455	2.19**	0.03**	ϕ_{12}	0.1929	0.1040	1.85*	0.06*
ϕ_{21}	-0.0345	0.0807	-0.43	0.67	ϕ_{21}	-0.0130	0.0301	-0.43	0.67
ϕ_{22}	-0.0925	0.0632	-1.46	0.14	ϕ_{22}	-0.1256	0.0624	-2.01**	0.04**

Panel B: $Sent_N$ **vs** *Return*

	Pre-transition Period					Post-transition Period			
ϕ_{11}	0.4469	0.0561	7.97***	0.00***	ϕ_{11}	0.4007	0.0572	7.00***	0.00***
ϕ_{12}	0.1060	0.0567	1.87*	0.06*	ϕ_{12}	0.2171	0.0949	2.29**	0.02**
ϕ_{21}	0.0849	0.0620	1.37	0.17	ϕ_{21}	0.0555	0.0374	1.48	0.14
ϕ_{22}	-0.0896	0.0626	-1.43	0.15	ϕ_{22}	-0.1257	0.0621	-2.02**	0.04**

Panel C: $Sent_S^2$ **vs** V_t

	Pre-transition Period					Post-transition Period			
ϕ_{11}	0.1520	0.0627	2.43**	0.02**	ϕ_{11}	0.5932	0.0508	11.67***	0.00***
ϕ_{12}	-0.0035	0.0677	-0.05	0.96	ϕ_{12}	-0.0386	0.1827	-0.21	0.83
ϕ_{21}	0.0270	0.0326	0.83	0.41	ϕ_{21}	-0.0027	0.0100	-0.27	0.79
ϕ_{22}	0.8327	0.0353	23.61***	0.00***	ϕ_{22}	0.8214	0.0358	22.95***	0.00***

Panel D: $Sent_N^2$ **vs** V_t

	Pre-transition Period					Post-transition Period			
ϕ_{11}	0.0385	0.0629	0.61	0.54	ϕ_{11}	0.3080	0.0607	5.07***	0.00***
ϕ_{12}	0.1889	0.1256	1.50	0.13	ϕ_{12}	0.6468	0.2877	2.25**	0.02**
ϕ_{21}	0.0089	0.0177	0.50	0.62	ϕ_{21}	0.0135	0.0078	1.74*	0.08*
ϕ_{22}	0.8287	0.0353	23.50***	0.00***	ϕ_{22}	0.8052	0.0368	21.90***	0.00***

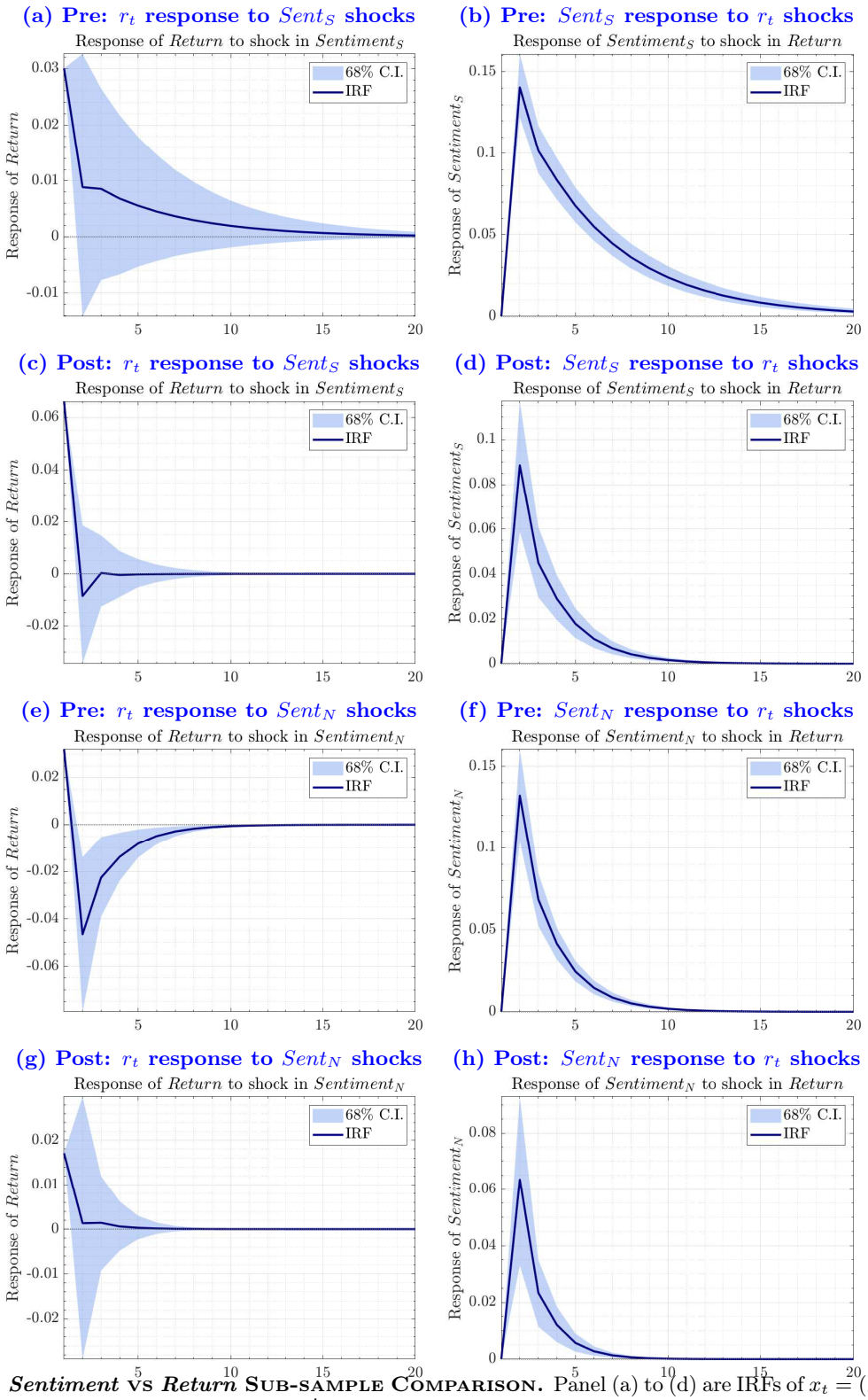


Figure 3: *Sentiment* vs *Return* Sub-sample Comparison. Panel (a) to (d) are IRFs of $x_t = (Sent_S, r_t)'$; panel (e) to (h) are IRFs of $x_t = (Sent_N, r_t)'$. "Pre" denotes Pre-transition period: 2011/01/01-2013/12/31; "Post" denotes Post-transition period: 2016/01/01-2017/11/30. Horizontal axis represent lagged days of IRFs. All time-series are standardized to have zero mean and unit variance. Error bands are constructed at the 68% interval following Sims and Zha (1999).

doubled the level after the transition. Second, the speed of revision for the temporary mispricing induced by social media sentiment has accelerated after 2016, comparing with that before 2014. In the pre-transition period, return reverts back to its original level in approximately 3 weeks (15 working days), while in the post-transition period, return shocks dissipate in only 2-3 days. Interestingly, the pattern of news media is just the opposite. The magnitude of initial impacts drops down from the pre-transition level of 0.030 (Panel (e)) to 0.016 in the post-transition period (Panel (g)) - approximately halved in value. Similar to the social media effects, the speed of reversion from news media influences also expedited in the post-transition period: return reverts back to its original level in about 8 to 9 working days in the pre-transition period (Panel (e)), but it only takes approximately 5 working days to revert in the post-transition period (Panel (g)).

Comparing Panels (a) and (e) in Figure 3, we find that, in the pre-transition period, returns are more sensitive to news sentiment impact than to social media sentiment. Panel (e) shows that with respect to a unit of shocks from news sentiment, returns over-correct to a negative level with a relatively narrower (more statistically significant) error band. In Panel (a), however, return gradually retreat with a wider error band with respect to shocks from social media sentiment. In contrast, a comparison between Panels (c) and (g) reveals that, in the post-transition period, returns exhibit strikingly higher sensitivity to social media sentiment impact than to news sentiment, as manifested by the higher initial reaction level (0.07 in Panel (c) vs 0.016 in Panel (g)) with a much narrower, thus more significant, error band in Panel (c) compared to Panel (g).

Panels of the IRFs on the right-hand side of Figure 3 indicate the reverse causalities of each of its respective left-hand side IRFs. All four panels (Panels (b), (d), (f) and (h)) exhibit similar patterns: a unit of shocks from stock return causes positive and significant increases in both social media based and news based sentiment the next day (observe spikes at lag 1 in the IRFs), and the increased sentiment revert back to zero exponentially at different speeds and in varied magnitudes. Similar to the results of the return responses, we find that the speed of sentiment reactions has also accelerated in the post-transition period. It takes about 20 working days for social media sentiment to correct itself before 2014 (Panel (b)), while it only takes approximately 12 working days to correct itself after 2016 (Panel (d)). Responses of news sentiment expedited, too. A unit of return shocks gives rise to rises in news sentiment that disappears in about 11 working days in the pre-transition period (Panel (f)), while this effect dies out in only approximately 7 working days in the post-transition sessions (Panel (h)).

Focusing on the magnitudes of sentiment responses (Panels (b), (d), (f) and (h) in Figure 3), we observe that both social media and news sentiment become less sensitive to returns at the post-transition period. For instance, a unit of return shocks results in 0.14 unit of heightened social media sentiment in the pre-transition period (Panel (b)), but this impact reduces to 0.09 unit in the post-transition period (Panel (d)). A unit of return shocks brings about 0.13 unit of news sentiment surges in the pre-transition session (Panel (f)), but this response contracts to a lower level of 0.065 at the post-transition stage (Panel (h)). It seems to be counter-intuitive to observe a reduced sensitivity to return in both social media and news sentiment (comparing Panels (b) with (d), and comparing (f) with (h)), but in fact it is not. One possible explanation to this phenomenon could be attributed to the scarcity of investor attention nowadays. The abundance

of communication platforms and information channels facilitates information exchanges among noise traders, but at the same time, it also dilutes individual tone or sentiment. As a result, a single opinion would be less influential under the increased information flow, leading to a lowered level of media sensitivity to stock return. Another feasible explanation for this decreased sensitivity might come from the stricter requirements from the censorship authority and regulatory bodies, as documented and exemplified in Section 1.

In sum, the findings of interaction between return and sentiment in this subsection validate and extend the media induced transition pattern identified in Section 4: social media effects become stronger after 2016, whereas news media plays the predominant role before 2014. For both return and sentiment series, the speeds of correction in IRFs with regard to unexpected shocks have accelerated in the post-transition period compared with the pre-transition period, irrespective of the types of media used in sentiment measure. Relative to the pre-transition period, the magnitude of return responses to social media sentiment have elevated in the post-transition period, while such magnitude dwindled with respect to news-based sentiment post-transition. Albeit stronger than the causal effects from sentiment to returns, feedback effects of returns on social on news media based sentiment have both depreciated in the post-transition period compared to the pre-transition levels.

5.2 Sentiment vs Volatility

Applying the same methodology in investigating the return-sentiment effects, we continue to explore the dynamic relationships between media sentiment and stock volatility at the pre- and post-transition periods. We estimate the following system equations, by representing $k = 2$, $x = (Sent_S^2, VIX)'$ and $x = (Sent_N^2, VIX)$ respectively into the General Setup.

$$\begin{aligned} Sent_{S,t}^2 &= \phi_{S,0} + \Phi_{1,1} Sent_{S,t-1}^2 + \Phi_{1,2} V_{t-1} + \epsilon_{1,t} \\ V_t &= \phi_{N,0} + \Phi_{2,1} Sent_{S,t-1}^2 + \Phi_{2,2} V_{t-1} + \epsilon_{2,t} \end{aligned} \tag{5}$$

$$\begin{aligned} Sent_{N,t}^2 &= \phi_{S,0} + \Phi_{1,1} Sent_{N,t-1}^2 + \Phi_{1,2} V_{t-1} + \epsilon_{1,t} \\ V_t &= \phi_{N,0} + \Phi_{2,1} Sent_{N,t-1}^2 + \Phi_{2,2} V_{t-1} + \epsilon_{2,t} \end{aligned} \tag{6}$$

We choose VIX (V_t) as a measure of volatility in the above two systems because investor sentiment affects asset prices by shaping investors' beliefs about the future. In contrast, traditional realized volatility measures (RV), such as standard deviation or squared term of prior returns, are backward-looking. Therefore, we believe that an implied, forward-looking volatility measure is more closely related to investor beliefs and more appropriate to this research. A detailed comparison between historical volatility and VIX is provided by Han and Park (2013). In order to assess whether VIX is associated with both positive and negative sentiment, we take the squared term of sentiment ($Sent_S^2$ and $Sent_N^2$) as a measure of the high sentiment period with strong emotions.[27] The benefit of using squared term of sentiment lies in its incorporation of

[27]The choice of $Sent^2$ is rooted in the fact that the relationship between volatility and sentiment is nonlinear. Only extreme values of sentiment show relationship with VIX, and both high negative and high positive sentiment resulted in high VIX values. The transformation is consistent with the choice of absolute value of sentiment, $|Sentiment|$, in Brown (1999, p.84, Hypothesis 1) when analysing correlation between sentiment and volatility. Berger and Turtle (2015, p.65, Eq.1) document quadratic relationship between sentiment and volatility of portfolio returns. Using maximum likelihood estimation of the Box-Cox power transformation parameter, λ, in a linear model of the form $Sent^\lambda = \alpha + \beta \times VIX$, we

the disagreement of opinions expressed in social and news media. Since our sentiment scores are volume-weighted[28] net values of positive and negative emotions conveyed in the parsed texts, the higher the $Sent^2$, the more likely that the grouped investors are driven by a similar kind of emotion. For example, when $Sent_S^2$ takes a value close to 1, most investors posting in social media are extremely optimistic, or are uniformly angry. Therefore, higher values of $Sent^2$ indicate less disagreement among investors' opinions. On the other hand, we interpret lower values of $Sent^2$ as containing more disagreement among investors' opinions, since a lower value of $Sent^2$ might result from: i) weak emotions expressed in media; and ii) strong positive and negative emotions expressed at the same time, but these parsed texts' scores cancelling with each other. We do not worry about this difference because both case indicate a higher level of disagreement of opinions. Similar to the return-sentiment mutual impacts analysed in prior subsection, we match TRMI sentiment data with VIX by averaging the non-trading days' sentiment indices, and standardise each variable to contain zero mean and unit standard deviation before importing each series to the VAR systems.

Panels C and D in Table 3 display the coefficients estimated and their level of significance for system equations (5) and (6) in the pre- and post-transition periods, respectively. These results suggest that the autocorrelation effect is more salient than the cross-impacts between sentiment and volatility. However, these values only indicate the initial responses, which do not help trace out the dynamics of responses for the dependent variable over time. Therefore, we put more emphasis on the impulse response functions (IRFs) rather than examining details of the VAR coefficients.

Left-hand side panels in Figure 4 (Panels (a), (c), (e) and (g)) depict the Impulse Response Functions (IRFs) of VIX responses to shocks from social media sentiment or news-based sentiment in both the pre- and post-transition periods. And the responses of media sentiment to shocks from VIX associated with the corresponding left panels, i.e. the feedback or reverse causality, are displayed in the right-hand side IRFs (Panels (b), (d), (f), and (h)). The top two panels in both sides (Panels (a), (b), (c), and (d)) are IRFs of the VIX and **social** media sentiment VAR system, while the bottom two panels in both sides (Panels (e), (f), (g) and (h)) are IRFs of the VIX and **news** media VAR system.

In both Panel (a) and (c) of Figure 4, we find that the extrema of VIX occur in 4 to 5 working days (about a week) following one unit of unexpected rise in social media sentiment, in the pre- and post-transition period respectively, and this process gradually corrects itself toward the original level. Error bands of these two IRFs do not cross zero, suggesting that volatility (VIX) responses are statistically different from zero over the IRFs forecasting window. In contrast to return responses (left side IRFs in Figure 3) that all revert back to zero within our IRFs observation window, the reaction of volatility (left side IRFs in Figure 4) dissipates after at least 20 working days (about a month), implying a more persistent effect compared to returns. In addition, we observe that in the pre-transition period, stock volatility is positively related to heightened social media sentiment (Panel (a)) - strong sentiment generates high VIX, while in the post-transition period (Panel (c)), volatility is negatively associated with the rising social

confirm our choice to square *Sentiment* variable at the 95% confidence level. Results are available upon request.

[28] Thomson Reuters MarketPsych Indices 2.2 User Guide, 23 March 2016, Document Version 1.0, Chapter 13, page 32: *'all emotional measures are "buzz-weighted" indices.'*

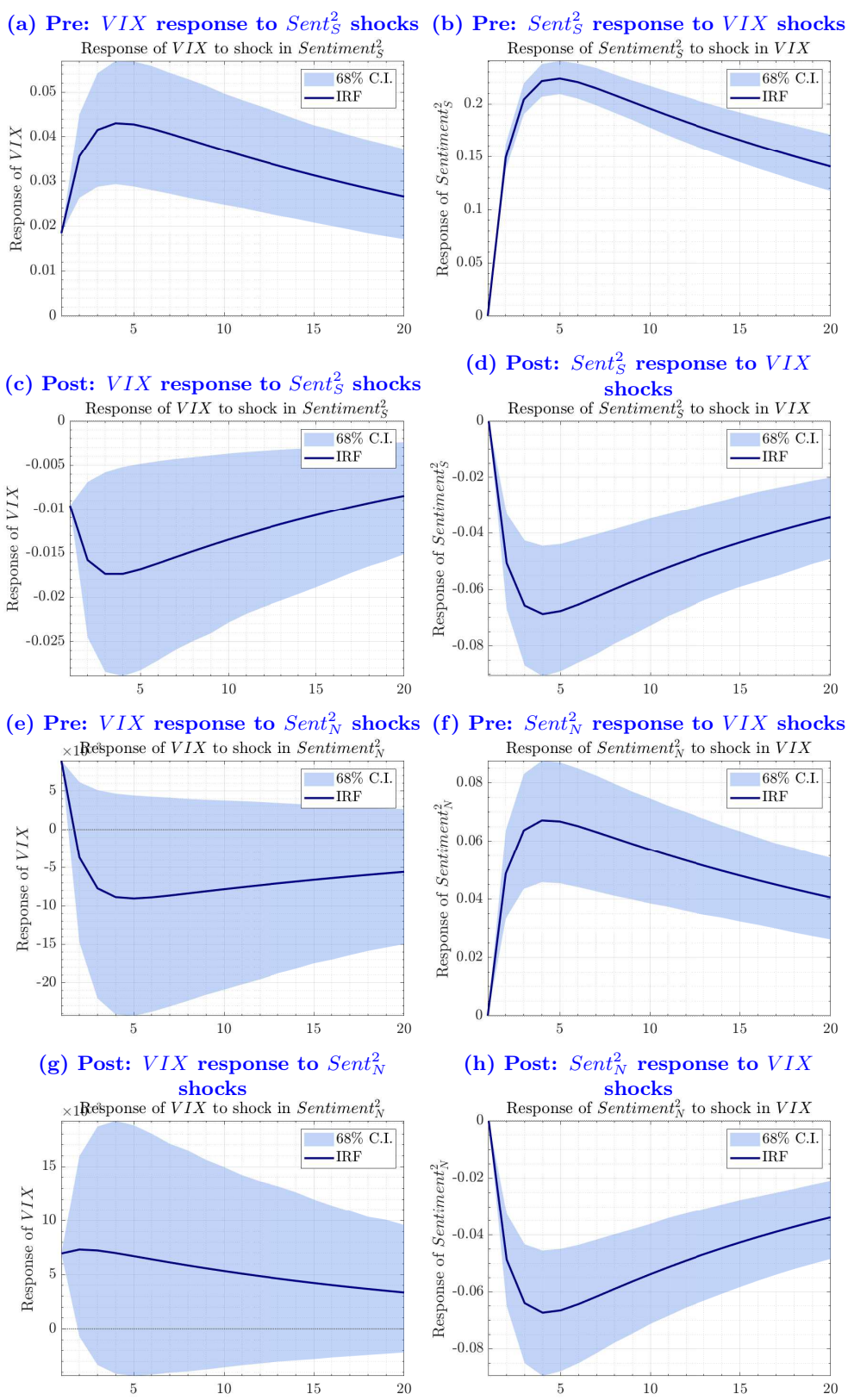


Figure 4: *Sentiment*^{2} **vs *VIX* Sub-sample Comparison.** Panels (a) to (d) are IRFs of $x_t = (Sent_S^2, V_t)'$; Panels (e) to (h) are IRFs of $x_t = (Sent_N^2, V_t)'$. "Pre" denotes Pre-transition Period: 2011/01/01-2013/12/31; "Post" denotes Post-transition Period: 2016/01/01-2017/11/30. Horizontal axis represents lagged days of IRFs (20 days). All time-series are standardized to have 0 mean and variance equal to 1. Error bands are constructed at the 68% interval following Sims and Zha (1999).

media sentiment. VIX responses to news sentiment shocks (Panels (e) and (g)), however, exhibit totally different patterns from that of social media. Comparing Panels (e) and (g), we find that initial VIX response to news sentiment shocks in the pre- and post-transition periods contain similar values (about 0.006 to 0.007). Interestingly, in Panel (e), the IRFs coefficients over-correct in 4 working days, whereas in Panel (g), the IRFs gradually dilute over the observation window. Consistent with the social media effects shown in Panels (a) and (c), estimated IRFs do not revert back to zero at the 20 working-day observation window. The broader error bands in Panels (e) and (g), which cross zero at lagged 1 to 2 days after the shock, indicate that volatility is less sensitive to news sentiment shocks than to social media sentiment shocks.

A comparison between the magnitudes of all four left-hand side panels with the right-hand side panels in Figure 4 reflects the fact that feedback effects from VIX to social media or news-based sentiment are stronger than the causal effects from media sentiment to VIX: the error bands of all four plots in the right side are significantly different from zero, and they are all narrower: statistically more significant than their left-hand side counterparts. In the pre-transition period, the positive IRFs in Panels (b) and (f) show that both social media and news-based sentiment spike higher following shocks from VIX, meaning the media formulates less disagreement with strong emotions after VIX surges higher. In the post-transition period, however, the upward concaved IRFs in Panels (d) and (h) illustrate that sentiment in social and news media with heightened VIX, regardless of the media type, becomes more neutral or contains more disagreement of opinion: both IFRs plots touch the troughs (approximately -0.07) after approximately 4 working days. In contrast to the fully correction situation in right side IRFs of Figure 3, none of the four right-hand side figures in Figure 4 displays fully correction after about 20 working days (one month), suggesting that VIX has a more persistent feedback effect on the media sentiment than return does.

To summarise the results in this section, although market volatility and media sentiment mutually cause each other, we find evidence that the feedback effects from market volatility on media sentiment prevail. This is consistent with Antweiler and Frank (2004), in their analysis of stock message boards, that stock messages help predict market volatility, however, the reverse feedback is stronger. Using TRMI data for the Brazilian market, Araújo et al. (2018) finds similar results, that is the reverse causal effects from VIX to social media or news-based sentiment are stronger than the causal effects from media sentiment to volatility. In addition to the aforementioned studies, we also find that VIX is more sensitive to social media sentiment than to news-based sentiment in terms of both reaction magnitudes and significance level. Comparing the pre- and post-transition periods, we observe that before 2014, high VIX level and strong emotions (or less disagreement) mutually cause each other, irrespective of the media type. After 2016, the heightened VIX is associated with neutral emotions (or more disagreement) for both social media and news. A comparison across the analysis performed for return in prior subsection (Figure 3) with analysis for volatility conducted in this subsection (Figure 4) reflects that, the mutual effects between media sentiment and volatility present a more persistent pattern than the inter-linkages between sentiment and return.

6 Conclusion

In this paper, we examine the dynamic relationships between social and news media activity, and the impact media has on the financial market. This paper contributes to the literature in several ways. Firstly, we examine the relationship between two different types of media, traditional news media and the rapidly growing social media. Our results show that the influence of news media in terms of generating activity and imparting sentiment have waned in the period between 2011 and 2017. By 2016, social media had become the dominant information source in generating media activity and sentiment. Next, we examine whether the rising influence of social media have permeated through the financial markets by examining the time-varying relationship between investor sentiment and the market. That is, we analyse causality between media sentiment and market variables (specifically, return and volatility) under different market information environments: (i) period of conventional news media dominance at the beginning of our sample, and (ii) period of increasing dominance of social media at the end of our sample. Our findings indicate that social media is becoming dominant. This should be of great interests (and possibly, concern) to regulators as social media is vulnerable to manipulation and misinformation.

We also discover that, generally, market variables exert stronger impact on investor sentiment than the other way around. That is, the reaction of media sentiment to stock market changes is more pronounced than the sensitivity of return and volatility to changes in media sentiment. However, when we contrast the two types of media at the pre- and post-transition periods, we find that return responses to social media sentiment almost doubled after the transition period, while the return responses to news-based sentiment almost halved to its pre-transition level. We observe that volatility in both pre- and post-transition periods display higher sensitivity to social media sentiment than to news-based sentiment. In addition, we find that the linkage between volatility and sentiment is much more persistent than that between returns and sentiment. These results corroborate our prior findings that social media is becoming the dominant media source. Overall, our exploration and results reveal new facts about the role of information in the social media era. An interesting extension in future work could focus on individual companies at a more granular frequencies to assess the timeliness of the two media types.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, 17(2):223–249.

Antweiler, W. and Frank, M. (2006). Do US stock markets typically overreact to corporate news stories? *Working Paper*.

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

Araújo, T., Eleutério, S., and Louçã, F. (2018). Do sentiments influence market dynamics? A reconstruction of the brazilian stock market and its mood. *Physica A: Statistical Mechanics and its Applications*, 505:1139–1149.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680.

Baker, M. and Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2):129–152.

Barber, B. M. and Odean, T. (2007). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21(2):785–818.

Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343.

Bartov, E., Faurel, L., and Mohanram, P. S. (2018). Can Twitter help predict firm-level earnings and stock returns? *The Accounting Review*, 93(3):25–57.

Berger, D. and Turtle, H. J. (2015). Sentiment bubbles. *Journal of Financial Markets*, 23:59–74.

Brown, G. W. (1999). Volatility, sentiment, and noise traders. *Financial Analysts Journal*, 55(2):82–90.

Brown, G. W. and Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 11(1):1–27.

Brzeszczyński, J., Gajdka, J., and Kutan, A. M. (2015). Investor response to public news, sentiment and institutional trading in emerging markets: A review. *International Review of Economics & Finance*, 40:338–352.

Canbaş, S. and Kandır, S. Y. (2009). Investor sentiment and stock returns: Evidence from Turkey. *Emerging Markets Finance and Trade*, 45(4):36–52.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403.

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461–1499.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *The Journal of Finance*, 53(6):1839–1885.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388.

De Bondt, W. F. and Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, 40(3):793–805.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 98(4):703–738.

DeMiguel, V., Nogales, F. J., and Uppal, R. (2014). Stock return serial dependence and out-of-sample portfolio performance. *The Review of Financial Studies*, 27(4):1031–1073.

Enders, W. (2014). *Applied Econometric Time Series*. Wiley Series in Probability and Statistics. Wiley.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements. *SSRN Electronic Journal*.

Engelberg, J. E., Reed, A. V., and Ringgenberg, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics*, 105(2):260–278.

Fang, L. and Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance*, 64(5):2023–2052.

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

Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100.

Gujarati, D. N. (2009). *Basic Econometrics*. Tata McGraw-Hill Education.

Han, H. and Park, M. D. (2013). Comparison of realized measure and implied volatility in forecasting volatility. *Journal of Forecasting*, 32(6):522–533.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195.

Heston, S. L. and Sinha, N. R. (2017). News vs sentiment: predicting stock returns from news stories. *Financial Analysts Journal*, 73(3):67–83.

Hirshleifer, D. and Teoh, S. H. (2009). Thought and behavior contagion in capital markets. In *Handbook of financial markets: Dynamics and evolution*, pages 1–56. Elsevier.

Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6):2143–2184.

Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.

Huang, R. D. and Stoll, H. R. (1997). The components of the bid-ask spread: A general approach. *The Review of Financial Studies*, 10(4):995–1034.

Ivanov, V. and Kilian, L. (2005). A practitioner's guide to lag order selection for var impulse response analysis. *Studies in Nonlinear Dynamics & Econometrics*, 9(1).

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

Jiao, P. and Walther, A. (2016). Social media, news media and the stock market. *SSRN Electronic Journal*.

Karlsson, N., Loewenstein, G., and Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, 38(2):95–115.

Kearney, C. and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web - WWW 2010*. ACM Press.

Lee, L. F., Hutton, A. P., and Shu, S. (2015). The role of social media in the capital market: Evidence from consumer product recalls. *Journal of Accounting Research*, 53(2):367–404.

Leung, H. and Ton, T. (2015). The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks. *Journal of Banking & Finance*, 55:37–55.

Liu, B. and McConnell, J. J. (2013). The role of the media in corporate governance: Do the media influence managers' capital allocation decisions? *Journal of Financial Economics*, 110(1):1–17.

Loughran, T. and McDonald, B. (2011a). Barron's red flags: Do they actually work? *Journal of Behavioral Finance*, 12(2):90–97.

Loughran, T. and McDonald, B. (2011b). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Michaelides, A., Milidonis, A., and Nishiotis, G. P. (2018). Private information in currency markets. *Journal of Financial Economics*.

Michaelides, A., Milidonis, A., Nishiotis, G. P., and Papakyriakou, P. (2015). The adverse effects of systematic leakage ahead of official sovereign debt rating announcements. *Journal of Financial Economics*, 116(3):526–547.

Nooijen, S. J. and Broda, S. A. (2016). Predicting equity markets with digital online media sentiment: Evidence from Markov-switching models. *Journal of Behavioral Finance*, 17(4):321–335.

Odean, T. (1999). Do investors trade too much? *American Economic Review*, 89(5):1279–1298.

Oliveira, N., Cortez, P., and Areal, N. (2013). On the predictability of stock market behavior using StockTwits sentiment and posting volume. In *Progress in Artificial Intelligence*, pages 355–365. Springer Berlin Heidelberg.

Peterson, R. (2013). Thomson Reuters Marketpsych Indices (TRMI) White Paper. *Inside the Mind of the Market*.

Peterson, R. (2016). *Trading on Sentiment: The Power of Minds Over Markets*. Wiley Finance Series. Wiley.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., and Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PloS one*, 10(9):e0138441.

Ren, Y. and Zhang, X. (2010). Subset selection for vector autoregressive processes via adaptive Lasso. *Statistics & Probability Letters*, 80(23-24):1705–1712.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking & Finance*, 84:25–40.

Sayim, M. and Rahman, H. (2015). The relationship between individual investor sentiment, stock return and volatility: evidence from the Turkish market. *International Journal of Emerging Markets*, 10(3):504–520.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shleifer, A. and Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, 52(1):35–55.

Siganos, A., Vagenas-Nanos, E., and Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization*, 107:730–743.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

Sims, C. A. and Zha, T. (1999). Error bands for impulse responses. *Econometrica*, 67(5):1113–1155.

Sprenger, T. O., Sandner, P. G., Tumasjan, A., and Welpe, I. M. (2014a). News or noise? Using Twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting*, 41(7-8):791–830.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014b). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957.

Stambaugh, R. F., Yu, J., and Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, 104(2):288–302.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.

Sun, L., Najand, M., and Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance*, 73:147–164.

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.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsay, R. S. (2005). *Analysis of Financial Time Series*, volume 543. John Wiley & Sons.

Wysocki, P. D. (1998). Cheap talk on the web: The determinants of postings on stock message boards. *Working paper 98025, University of Michigan Business School*.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

A Appendix

A.1 List of acronyms and notation

Table A.1: List of acronyms, data sources and variable names.

Acronym	Description
AAII	American Association of Individual Investors
ACF	Autocorrelation Function
AIC	Akaike Information Criterion
BIC	Schwartz's Bayesian Information Criterion
BW	Baker & Wurgler sentiment index
BW_O	The orthoganolized Baker & Wurgler sentiment index
CBOE	Chicago Board Options Exchange
CEFD	Closed-end fund discount
Datastream	Thomson Reuters Datastream
DJIA	Dow Jones Industry Average
DJNS	Dow Jones Newswires
DW	Durbin-Watson test
GFC	Global Financial Crisis
GI	Harvard General Inquirer Dictionary
GSV	Google Search Volume
IQR	Interquartile Range
IRF	Impulse Response Function
LB	Ljung-Box test
MV	Market Variables
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
RIC	Reuters Identification Code
S&P500	Standard & Poor's 500 Index
SEC	The US Securities and Exchange Commission
SIRCA	Securities Industry Research Centre of Asia-Pacific
VAR	Vector Autoregressive Model
TR	Thomson Reuters
TRMI	Thomson Reuters MarketPsych Indices
TRNA	Thomson Reuters News Analytics
TRNS	Thomson Reuters News Scope
TRTH	Thomson Reuters Tick History
VAR	Vector Autoregressive Model
WRDS	Wharton Research Data Services
WSJ	The Wall Street Journal

Code/Symbol	Description
.SPY	RIC for SPDR ETF
Datastream	Thomson Reuters Datastream
MPTRXUS500	TRMI company group code approximating S&P500 index constituents
$Buzz_{N,t}$	(N)ews media buzz at time t (report volume in news media)
$Buzz_{S,t}$	(S)ocial media buzz at time t (posting volume in social media)
$Sent_{N,t}$	(N)ews media net sentiment at time t (positive minus negative sentiment)
$Sent_{S,t}$	(S)ocial media net sentiment at time t (positive minus negative sentiment)
r_t	log return on day t
V_t	VIX (CBOE options volatility index) on day t

A.2 Testing for Unit Roots

Prior to fitting VAR models, we perform tests for unit root to ensure that all regressors are covariance stationary. Results in Table A.2 suggest rejection of the null hypothesis of unit root for all series.

Table A.2: Unit Root Test. The table displays Augmented Dickey-Fuller and Phillips-Perron unit root test statistics for 3 different model variants. ***, **, and * denote rejection of the null hypothesis of unit-root at 1%, 5%, and 10% significance levels, respectively. Critical values at 5% significance level are reported in the last column.

	Return	VIX	$Sent_S$	$Sent_N$	$Buzz_S$	$Buzz_N$	Crit.Val.
Panel A: Augmented Dickey-Fuller (ADF) test							
ADF	-44.76***	-1.83*	-11.12***	-17.11***	-3.02***	-3.03***	-1.94
ADF (with drift)	-44.85***	-5.45***	-13.93***	-19.63***	-11.46***	-15.04***	-2.86
ADF (with drift and trend)	-44.84***	-6.36***	-17.95***	-19.78***	-12.93***	-15.04***	-3.41
Panel B: Phillips-Perron (PP) test							
PP	-29.47***	-1.74*	-8.88***	-13.41***	-2.76***	-2.85***	-1.94
PP (with drift)	-29.56***	-5.16***	-10.92***	-15.36***	-10.28***	-14.17***	-2.86
PP (with drift and trend)	-29.56***	-6.03***	-13.82***	-15.47***	-11.61***	-14.17***	-3.41

A.3 ACF and PACF for main TRMI series

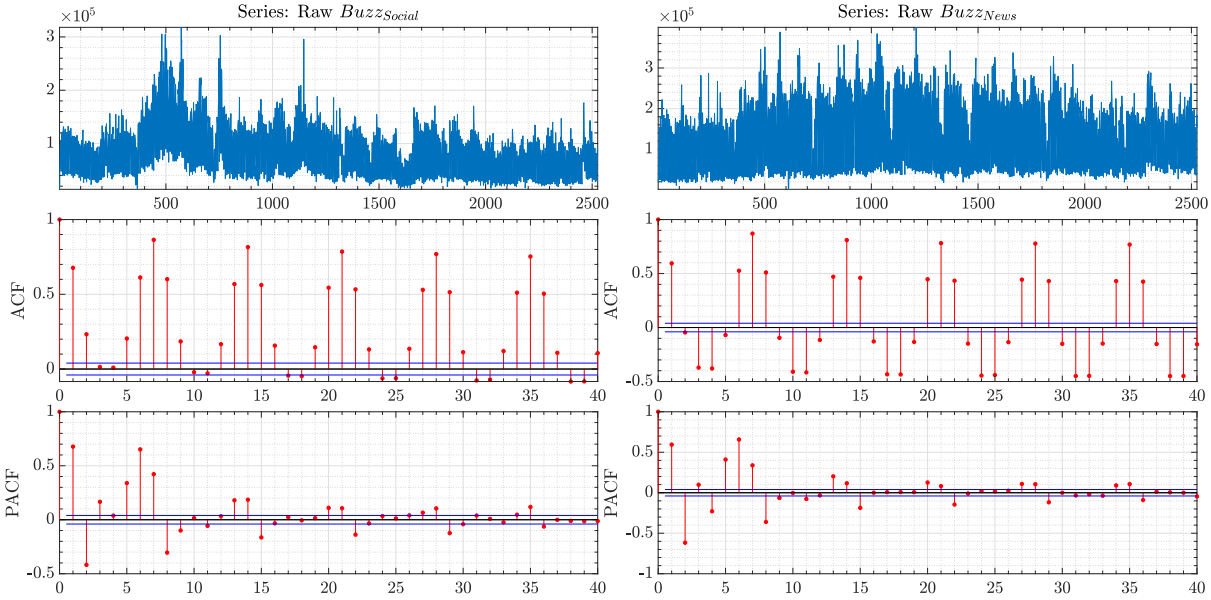


Figure A.1: Time-series Analysis of raw *Buzz* data. The left three panels show the sample distribution of the original social media posts volume measure: $Buzz$, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF) up to 40 days. The three panels on the right represent news-based $Buzz$ series distribution, its ACF and PACF respectively. Sampling period: 2011/01/01-2017/11/30. The top two figures (blue series) verify descriptive statistics reported in Table 1, and highlight the fact that the original $Buzz$ series contain several observations at the right tail (large outliers). Social (left) $Buzz$ tends to be more volatile than news (right) counterpart. Both ACF and PACF indicate the presence of strong weekly seasonality for both $Buzz_S$ abd $Buzz_N$

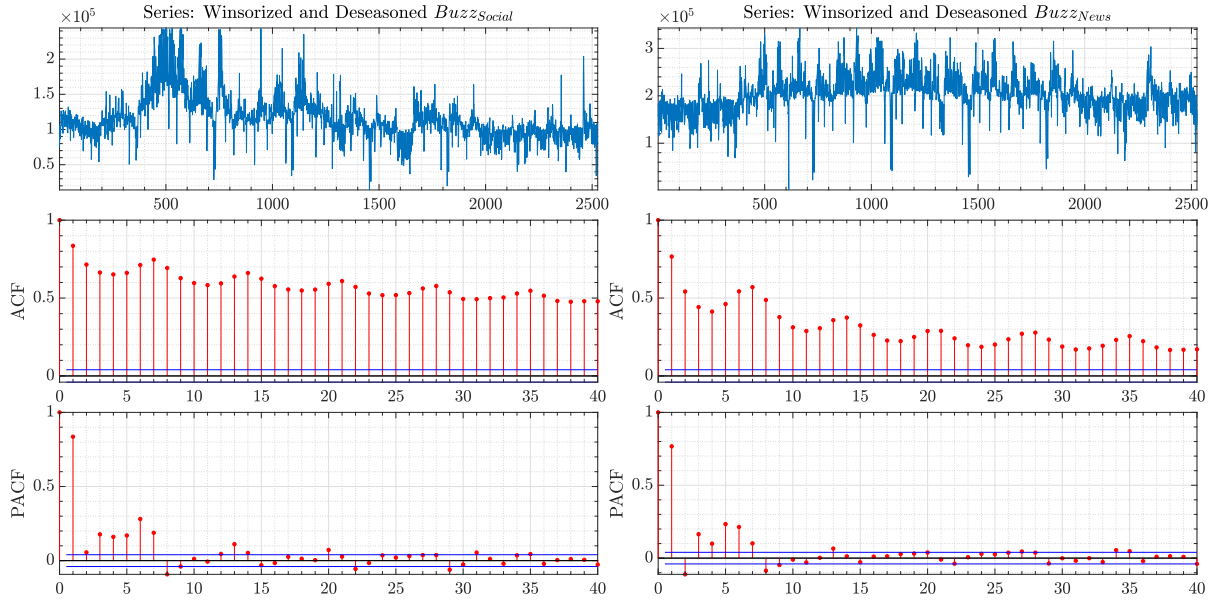


Figure A.2: Winsorized and De-Seasoned *Buzz* Time Series Check. The left three panels show the sample distribution of $Buzz_S$ after truncating the large value observations (asymmetric winsorizing the right tail outliers), its autocorrelation function (ACF) and partial autocorrelation function (PACF) up to 40 days. The right side three panels represent the winsorized and seasonality adjusted news-based $Buzz$, its ACF and PACF respectively. Sampling period: 2011/01/01-2017/11/30. Comparing with Figure A.1, the ACFs and PACFs of these two series indicate a diminished, yet not fully eliminated weekly seasonality. Since this research does not involve the association between $Buzz$ and stock returns/volatility, the non-trading day adjusted $Buzz$ distributions are not reported for brevity.

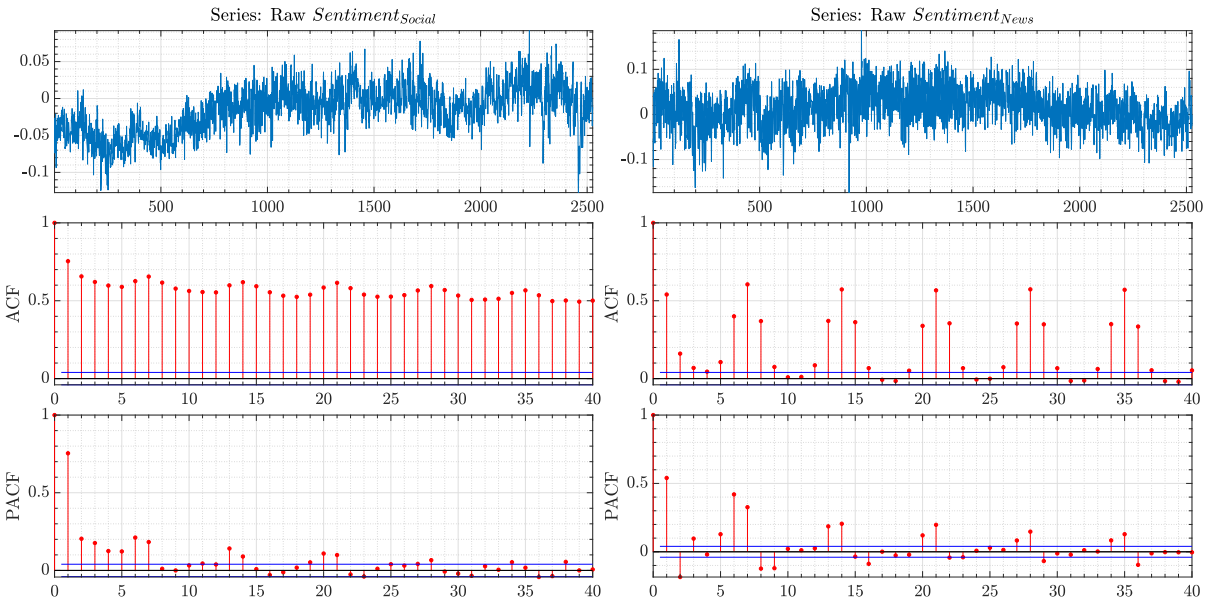


Figure A.3: Raw *Sentiment* Time Series Check. The left three panels show the sample distribution of the net positive and negative emotion scores from social media: $Sent_S$, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF) up to 40 days. The right side three panels represent news-based *Sentiment* series distribution, its ACF and PACF respectively. Sampling period: 2011/01/01-2017/11/30. The top two figures (blue series) illustrate that the original *Sentiment* series are normalised to zero mean, consistent with descriptive statistics from Table 1. Social (left) *Sentiment* exposes more negative observations than news-based (right) scores. Both ACF and PACF indicate the existence of weekly seasonality, and this property is more obvious in news-based sentiment scores.

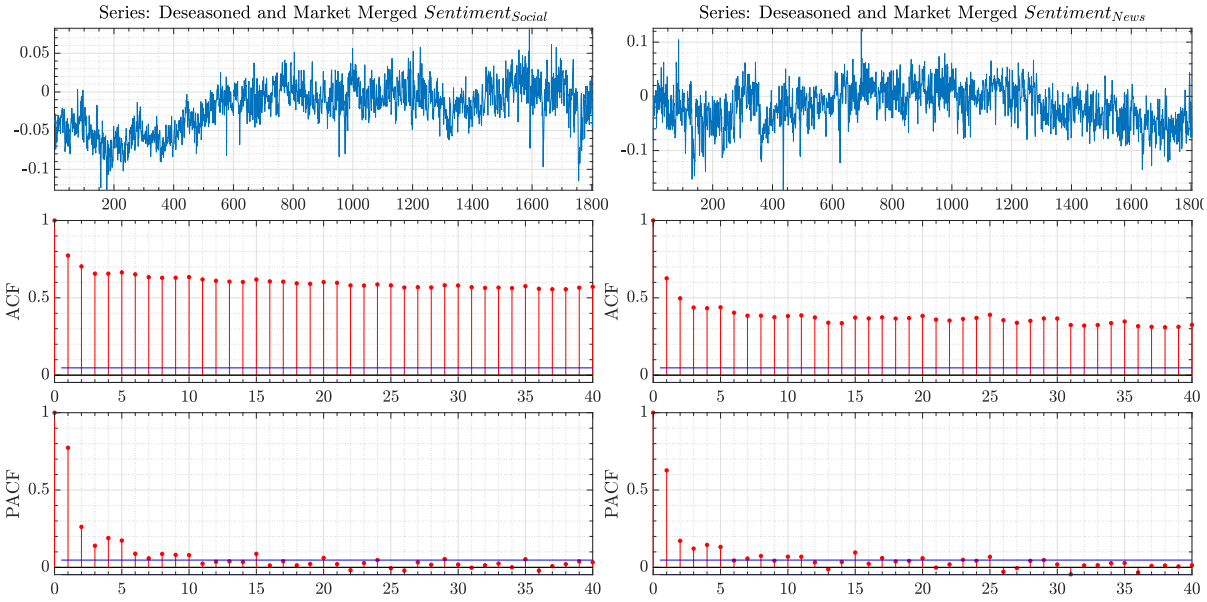


Figure A.4: De-Seasoned and Market Merged *Sentiment* Time Series Check. The left three panels show the sample distribution of the seasonality adjusted and non-trading day averaged value of $Sent_S$, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF) up to 40 days. The right side three panels represent news-based *Sentiment* series distribution after dealing with the weekly effects and merging with the trading-day only market variables. Its ACF and PACF are presented below respectively. Sampling period: 2011/01/01-2017/11/30. Since *Sentiment* are volume ($Buzz$) weighted and normalised, we do not winsorize *Sentiment* series. This research concentrates on the inter-relations between *Sentiment* and stock variables, we match the *Sentiment* scores with market variables by averaging the non-trading day values. Both ACF and PACF indicate that the weekly seasonality is properly tackled with after these procedures.

B Supplementary Online Appendix

B.1 Tried-and-true vs Bold-and-New: on commonality between Baker & Wurgler and MarketPsych Indices

Recently launched Thomson Reuters MarketPsych Indices (TRMI) contain synthesized quantities and emotional measures from a wide range of traditional news channels as well as social media platforms. We contrast sentiment captured by TRMI from social and news media to the "tried-and-true" Baker & Wurgler index (BW) commonly used in investor sentiment analysis in the past decade. To do this, we aggregate the daily TRMI social media and news sentiment scores (denoted as $Sent_S$ and $Sent_N$ respectively) into monthly frequency and report the correlations between TRMI and the BW sentiment indices in Table B.1. The results in Table B.1 demonstrate commonalities between TRMI sentiment indicators and the BW index, yet, the magnitude of correlation coefficients are indicative of divergence of these two measures. This suggests that the TRMI sentiment indices capture different investor sentiment from BW's. Thus, on one hand, strong positive correlation provides merit for using TRMI as it captures commonality in general trend of these two indicators. On the other hand, TRMI provides sentiment scores at a much higher frequencies allowing us to study the dynamics in temporal displacement within sentiment scores (news vs social) and between sentiment and market variables (sentiment vs returns and/or volatility).

Table B.1: Correlation Between BW and TRMI Sentiment Indices. Sample period Jan/2011-Sep/2015. TRMI daily sentiment indices are aggregated into monthly frequency to match the BW index. BW sentiment data are obtained from personal website of Jeffrey Wurgler at NYU Stern. BW and BW_O denote the investor sentiment from equation (2) and the orthogonalized sentiment index from equation (3) of Baker and Wurgler (2006) respectively. ***, **, and * indicate significance levels of 1%, 5%, and 10% respectively.

	$Sent_S$	$Sent_N$	BW	BW_O
$Sent_S$	1.000			
$Sent_N$	0.784***	1.000		
BW	**0.543*** **	**0.440*** **	1.000	
BW_O	**-0.358*** **	**-0.318** **	0.339***	1.000

B.2 Robustness check: model selection

Model selection is an integral part of the statistical analysis of VAR models. For VAR models, model selection consists of two parts:

1. determining the lag order, and
2. determining the substructures of the VAR model.

Much of the existing literature on VAR model selection focus only on the first part, i.e., the lag order determination part, presumably because that misspecification of the lag order often has undesirable implications for subsequent analysis.

When the selected lag order is underfitted, there can be significant residual autocorrelations. Simulations of Ivanov and Kilian (2005) revealed that lag order selection is practically important for impulse response analysis. A number of approaches have been proposed for lag order selection, including the information criterion based approaches such as AIC (Akaike, 1998), BIC (Schwarz, 1978) and HQC (Hannan and Quinn, 1979), the hypothesis testing based approaches such as the sequential likelihood ratio test.

Recently, with the development of penalty-based variable selection techniques such as the Lasso (Tibshirani, 1996) and the adaptive Lasso (Zou, 2006), researchers have begun to consider both parts of VAR model selection simultaneously. Hsu et al. (2008) applied the idea of the Lasso to VAR models to select the lag order and determine the substructures of the coefficient matrices all together. Ren and Zhang (2010) proposed a model selection method using the adaptive Lasso. Although most of the above-mentioned methods have solid theoretical justifications, simulation study results are mixed and usually conflicting, and a universally acceptable method is still unavailable.

In what follows, we present our results from AIC and BIC estimation in Subsection B.2.1. We detail and contrast estimates of bivariate VAR(1) and VAR(7) systems and discuss why we prefer a more parsimonious VAR(1) system in Subsection B.2.2. We perform a formal likelihood ratio test by sequentially contrasting VAR($p-1$) vs VAR(p) models. We present our most conservative test results when comparing the most restrictive VAR(1) model to the least restrictive VAR(7) in Subsection B.2.3. Faced with potential omitted variable bias in our estimation of off-diagonal elements in bivariate VARs, we check the robustness of these coefficients by estimating VAR systems with expanded set of variables. Subsection B.2.4 details our findings.

B.2.1 Optimal Lag Length: Information Criterion

One of most frequent approach to specification is the use of information criteria. The logic behind is the following: we want to minimize the sum of squared error, but in the meantime we want to penalize the dimension of the model (and so the loss of degrees of freedom). Note that the BIC criterion puts a stronger penalty on the number of regressors. Choose the model with the smallest BIC.

Table B.2: Optimal Lag Selection using Information Criteria. Panel A tabulates AIC and BIC criteria from lag 1 to lag 12 for the VAR systems between social media *Buzz* and news *Buzz*. Similarly, Panel B lists AIC and BIC values for the VAR with social media *Sentiment* and news *Sentiment*. Optimal lag is denoted with * and boldface. BIC selects more parsimonious models by imposing heavier penalties for number of lags. AIC is included to facilitate judgment and for completeness. BICs of Panels A and B suggest that the optimal lag for investigating social media and news dynamics are 7. Likewise, Panels C and D test optimal lags for the VAR systems between *Sentiment* and *Return* for social and news media respectively, 5 lags is detected to be most suitable. Panels E and F list the AIC and BIC of VARs between $Sentiment^2$ and VIX for social and news media respectively, BIC indicates that 2 lags are most appropriate for the model specification.

Panel A: $Buzz_S$ **vs** $Buzz_N$

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Lag 11	Lag 12
AIC	3.189	3.153	3.092	3.055	2.988	2.847	2.787	2.773	**2.772***	2.776	2.777	2.776
BIC	3.203	3.176	3.125	3.097	3.039	2.907	2.856	**2.852***	2.861	2.873	2.884	2.892

Panel B: $Sent_S$ **vs** $Sent_N$

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Lag 11	Lag 12
AIC	4.165	4.079	4.031	4.005	3.982	3.941	3.911	**3.909***	3.911	3.912	3.913	3.911
BIC	4.179	4.102	4.064	4.047	4.033	4.001	**3.981***	3.988	3.999	4.01	4.019	4.027

Panel C: $Sent_S$ **vs** *Return*

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Lag 11	Lag 12
AIC	4.735	4.655	4.624	4.588	4.549	4.545	4.543	4.541	4.536	**4.534***	4.537	4.541
BIC	4.754	4.686	4.667	4.643	**4.616***	4.625	4.635	4.645	4.653	4.663	4.679	4.695

Panel D: $Sent_N$ **vs** *Return*

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Lag 11	Lag 12
AIC	5.159	5.122	5.098	5.082	**5.06***	5.063	5.064	5.065	5.067	5.063	5.062	5.063
BIC	5.177	5.153	5.141	5.138	**5.128***	5.143	5.156	5.169	5.183	5.193	5.204	5.217

Panel E: $Sent_S^2$ **vs** VIX

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Lag 11	Lag 12
AIC	3.229	3.203	3.197	3.176	3.169	3.159	3.160	**3.148***	3.148	3.149	3.152	3.152
BIC	3.248	**3.234***	3.240	3.231	3.236	3.238	3.252	3.252	3.264	3.278	3.293	3.305

Panel F: $Sent_N^2$ **vs** VIX

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Lag 11	Lag 12
AIC	3.706	3.688	3.693	3.680	3.681	3.678	3.682	**3.674***	3.677	3.679	3.683	3.686
BIC	3.724	**3.719***	3.735	3.735	3.748	3.758	3.774	3.779	3.794	3.808	3.825	3.840

B.2.2 Optimal Lag Length: Why VAR(1) is Parsimonious Form of VAR (7)

Table B.3: VAR(7) Parsimonious Form Examination (A). Sample A: 2011/01/01-2011/12/31 (the **first year** of our sampling period); p-values smaller than 0.1, 0.05 and 0.01 are denoted as *, **, and *** respectively. Left panel shows VAR model coefficients estimated as in the General Setup when $\mathbf{x} = (Buzz_S, Buzz_N)'$ and $p = 7$; right panel indicates coefficients estimated when $\mathbf{x} = (Sent_S, Sent_N)'$ and $p = 7$. The p-value columns show that the innner lags' (lag 2 to lag 6's) coefficients are insignificant in both models, and most of the significant coefficients are concentrated on lag 1 and lag 7. This indicates that VAR(1) might be a parsimonious form representation of VAR(7).

Sample A: First 365 days							
VAR(7): $Buzz_S$ **vs** $Buzz_N$				**VAR(7):** $Sent_S$ **vs** $Sent_N$			
	Coef.	*s.e.*	p-value		Coef.	*s.e.*	p-value
Constant1	-0.0742	0.0460	0.11	Constant1	-0.2109	0.0724	0.00***
Constant2	-0.2417	0.0694	0.00***	Constant2	0.0884	0.1284	0.49
AR1(1,1)	0.6312	0.0683	0.00***	AR1(1,1)	0.5072	0.0589	0.00***
AR1(2,1)	0.0198	0.1030	0.85	AR1(2,1)	0.2718	0.1044	0.01***
AR1(1,2)	-0.0202	0.0452	0.65	AR1(1,2)	-0.0283	0.0337	0.40
AR1(2,2)	0.6084	0.0681	0.00***	AR1(2,2)	0.3482	0.0597	0.00***
AR2(1,1)	-0.0046	0.0802	0.95	AR2(1,1)	-0.0243	0.0648	0.71
AR2(2,1)	0.0954	0.1209	0.43	AR2(2,1)	-0.0939	0.1149	0.41
AR2(1,2)	-0.0687	0.0525	0.19	AR2(1,2)	-0.0028	0.0361	0.94
AR2(2,2)	-0.2647	0.0792	0.00***	AR2(2,2)	0.0487	0.0639	0.45
AR3(1,1)	0.0336	0.0803	0.68	AR3(1,1)	0.1230	0.0645	0.06
AR3(2,1)	-0.0353	0.1210	0.77	AR3(2,1)	-0.0374	0.1143	0.74
AR3(1,2)	0.0228	0.0534	0.67	AR3(1,2)	-0.0387	0.0361	0.28
AR3(2,2)	0.0878	0.0806	0.28	AR3(2,2)	0.0748	0.0640	0.24
AR4(1,1)	-0.0306	0.0804	0.70	AR4(1,1)	0.0238	0.0650	0.71
AR4(2,1)	-0.0955	0.1212	0.43	AR4(2,1)	-0.0425	0.1151	0.71
AR4(1,2)	-0.0153	0.0535	0.78	AR4(1,2)	0.0245	0.0362	0.50
AR4(2,2)	-0.0257	0.0806	0.75	AR4(2,2)	0.0891	0.0642	0.17
AR5(1,1)	0.0810	0.0805	0.31	AR5(1,1)	0.1037	0.0642	0.11
AR5(2,1)	0.1523	0.1213	0.21	AR5(2,1)	-0.0070	0.1138	0.95
AR5(1,2)	-0.0537	0.0534	0.31	AR5(1,2)	0.0042	0.0361	0.91
AR5(2,2)	-0.1052	0.0805	0.19	AR5(2,2)	0.0134	0.0640	0.83
AR6(1,1)	0.0876	0.0812	0.28	AR6(1,1)	0.0558	0.0643	0.38
AR6(2,1)	-0.0882	0.1223	0.47	AR6(2,1)	0.0662	0.1139	0.56
AR6(1,2)	0.0189	0.0527	0.72	AR6(1,2)	-0.0299	0.0360	0.41
AR6(2,2)	0.2426	0.0795	0.00***	AR6(2,2)	-0.0062	0.0638	0.92
AR7(1,1)	0.0142	0.0686	0.84	AR7(1,1)	0.0390	0.0591	0.51
AR7(2,1)	-0.1398	0.1034	0.18	AR7(2,1)	0.0783	0.1047	0.45
AR7(1,2)	0.0876	0.0455	0.05**	AR7(1,2)	0.0725	0.0334	0.03**
AR7(2,2)	0.2093	0.0686	0.00***	AR7(2,2)	0.0230	0.0592	0.70

[continue table next page]

Table B.4: VAR(7) Parsimonious Form Examination (B). Sample B: 2016/11/30-2017/11/30 (the **last year** of our sampling period); p-values smaller than 0.1, 0.05 and 0.01 are denoted as *, **, and *** respectively. Left panel shows VAR model coefficients estimated as in the General Setup when $\mathbf{x} = (Buzz_S, Buzz_N)'$ and $p = 7$; right panel expresses coefficients estimated when $\mathbf{x} = (Sent_S, Sent_N)'$ and $p = 7$. The results indicate that the inner lags' (lag 2 to lag 6's) coefficients are insignificant in both models, and most of the significant coefficients are concentrated on lag 1 and lag 7. This indicates that VAR(1) might be a parsimonious form representation of VAR(7).

Sample B: Last 365 days

VAR(7): $Buzz_S$ **vs** $Buzz_N$				**VAR(7):** $Sent_S$ **vs** $Sent_N$			
	Coef.	*s.e.*	p-value		Coef.	*s.e.*	p-value
Constant1	-0.1695	0.0451	0.00***	Constant1	-0.0039	0.0959	0.97
Constant2	-0.0429	0.0563	0.45	Constant2	-0.3800	0.0916	0.00***
AR1(1,1)	0.6037	0.0680	0.00***	AR1(1,1)	0.6260	0.0560	0.00***
AR1(2,1)	-0.0516	0.0848	0.54	AR1(2,1)	0.1098	0.0535	0.04**
AR1(1,2)	0.0462	0.0549	0.40	AR1(1,2)	-0.0809	0.0595	0.17
AR1(2,2)	0.7532	0.0686	0.00***	AR1(2,2)	0.4117	0.0568	0.00***
AR2(1,1)	0.0029	0.0804	0.97	AR2(1,1)	0.0107	0.0654	0.87
AR2(2,1)	-0.0422	0.1004	0.67	AR2(2,1)	-0.0727	0.0624	0.24
AR2(1,2)	-0.1293	0.0675	0.06	AR2(1,2)	0.0049	0.0651	0.94
AR2(2,2)	-0.2267	0.0842	0.01***	AR2(2,2)	0.0646	0.0622	0.30
AR3(1,1)	-0.0200	0.0802	0.80	AR3(1,1)	-0.0844	0.0655	0.20
AR3(2,1)	-0.0205	0.1002	0.84	AR3(2,1)	0.0203	0.0625	0.75
AR3(1,2)	0.0768	0.0679	0.26	AR3(1,2)	0.0177	0.0651	0.79
AR3(2,2)	0.1312	0.0848	0.12	AR3(2,2)	-0.0547	0.0621	0.38
AR4(1,1)	-0.0323	0.0802	0.69	AR4(1,1)	0.0705	0.0653	0.28
AR4(2,1)	-0.0059	0.1001	0.95	AR4(2,1)	-0.0980	0.0623	0.12
AR4(1,2)	-0.0253	0.0681	0.71	AR4(1,2)	-0.0206	0.0649	0.75
AR4(2,2)	-0.0499	0.0851	0.56	AR4(2,2)	0.0144	0.0620	0.82
AR5(1,1)	-0.0071	0.0802	0.93	AR5(1,1)	0.0625	0.0654	0.34
AR5(2,1)	0.0427	0.1002	0.67	AR5(2,1)	0.0984	0.0624	0.12
AR5(1,2)	-0.0319	0.0681	0.64	AR5(1,2)	-0.0065	0.0648	0.92
AR5(2,2)	-0.0019	0.0850	0.98	AR5(2,2)	-0.0354	0.0619	0.57
AR6(1,1)	0.1152	0.0802	0.15	AR6(1,1)	-0.0596	0.0658	0.36
AR6(2,1)	0.0482	0.1001	0.63	AR6(2,1)	-0.0121	0.0628	0.85
AR6(1,2)	0.0334	0.0673	0.62	AR6(1,2)	-0.0223	0.0649	0.73
AR6(2,2)	0.1582	0.0840	0.06	AR6(2,2)	0.0382	0.0620	0.54
AR7(1,1)	0.0404	0.0685	0.55	AR7(1,1)	0.2191	0.0568	0.00***
AR7(2,1)	0.0800	0.0855	0.35	AR7(2,1)	0.0262	0.0542	0.63
AR7(1,2)	0.0742	0.0543	0.17	AR7(1,2)	-0.0091	0.0596	0.88
AR7(2,2)	0.0512	0.0678	0.45	AR7(2,2)	0.1243	0.0569	0.03**

B.2.3 Optimal Lag Length: Likelihood Ratio test

Appropriate lag length selection can be critical. In this section, we investigate the appropriateness of our lag choice. If the number of lags in VAR system is too small, the model is misspecified; if the number of lags is too large, degrees of freedom are wasted. The likelihood ratio test, which evaluates the statistical significance of the difference in log-likelihoods at the unrestricted and restricted parameter estimates, is generally considered to be the most reliable of the three classical tests of model specification (namely, Likelihood Ratio test, Wald test, and Lagrange Multiplier test).

We reconfirm our selection of VAR(1) over VAR(7) with likelihood ratio test. Our results are presented in Figure B.1. Our goal is to determine whether bivariate VAR systems, $Buzz_S$ and $Buzz_N$ (top panel) and $Sent_S$ and $Sent_N$ (bottom panel), containing only one lag are indeed appropriate. The proper test for this cross-equation restriction is a likelihood ratio test.[29] Given the sample size restriction in our rolling window analysis, we follow recommendations in Sims (1980) and compute the likelihood ratio statistic as

$$(T - c)\left(\ln|\Sigma_1| - \ln|\Sigma_7|\right),$$

where T is number of observations, c the number of parameters estimated in each equation of the unrestricted system, Σ_n the covariance matrix of residuals from VAR(n) system, and $\ln|\Sigma_n|$ the natural logarithm of the determinant of Σ_n. This statistic has an asymptotic χ^2_k distribution with degrees of freedom, k, equal to the number of restrictions in the system.

Using estimation window of 252 days (for consistency with our main analysis), we perform a rolling window VAR estimation on each day in our sample. The number of estimated VAR systems is 1,546 for each series set for each lag length. p-values of the likelihood ratio test along time using the equation above are plotted in Figure B.1. At the beginning, mid-sample, and at the end of our sample period, VAR(1) systems are appropriate. There are two sub-periods, namely 2013 - 2014 and 2016, where estimation would have benefited from VAR systems allowing for larger number of lags. Since our objective is to contrast the earlier period to the later period, our decision in selecting VAR(1) model is justified for both *Buzz* and *Sentiment* series.

[29] We followed the procedure outlined in Enders (2014, pp.303-305).

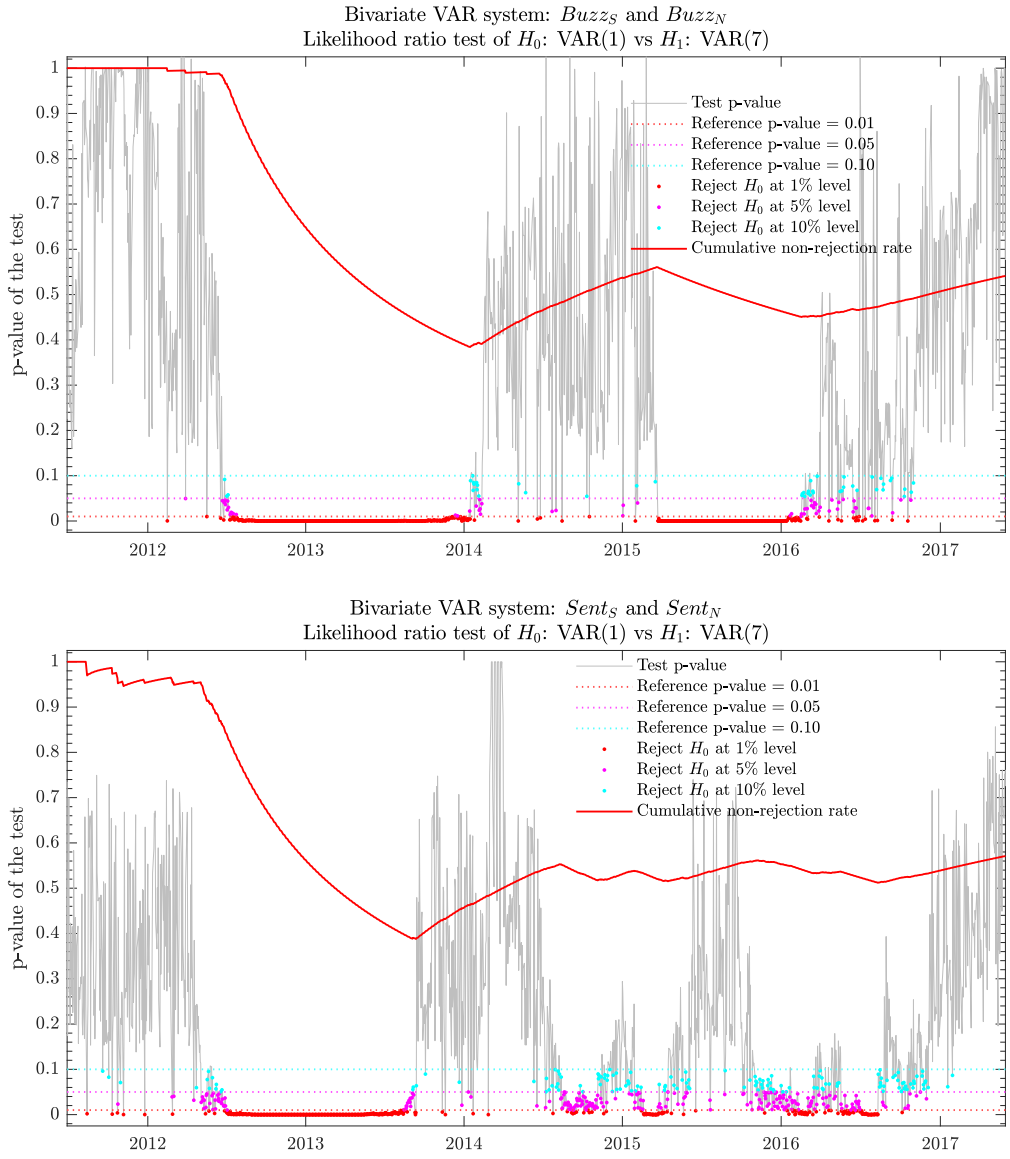


Figure B.1: Likelihood ratio test results The figure displays p-values overtime from the likelihood ratio test by contrasting the test statistic, $(T-c)\left(\ln|\Sigma_1| - \ln|\Sigma_7|\right)$ to critical values of χ^2 distribution.

B.2.4 Model specification: VAR subsystems and omitted variable bias

As pointed out in Stock and Watson (2001), the VAR shocks, like those in conventional regression, reflect factors omitted from the model. If these factors are correlated with the included variables, then the VAR estimates will contain omitted variable bias ensuing undesirable implications for subsequent analysis.

In contrast to Figure 1, where results are depicted for the bivariate system, $\mathbf{x}_t = (Buzz_S, Buzz_N)'$, in Figure B.2 we present results from the four-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Buzz_S, Buzz_N)'$ in the top panel and a six-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Buzz_S, Sent_S, Buzz_N, Sent_N)'$ in the bottom panel. The pattern in the lead-lag dynamics between social media buzz and news media buzz is strikingly similar. Even with inclusion of 4 additional variables, the change in estimated coefficients is minimal. More importantly, the sign and significance of the estimates is still in accordance with the bivariate VAR(1) system in Figure 1. Given the sample size restrictions and the degrees-of-freedom constraints, we allude to the simpler bivariate form VAR(1) model as the best fit.

Similarly, to contrast Figure 2, where results are depicted for the bivariate system, $\mathbf{x}_t = (Sent_S, Sent_N)'$, in Figure B.3 we present results from the four-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Sent_S, Sent_N)'$ in the top panel and a six-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Buzz_S, Sent_S, Buzz_N, Sent_N)'$ in the bottom panel. The pattern in the lead-lag dynamics has larger deviations compared to $Buzz$-focused systems as discussed in previous paragraph. Nevertheless, similarity among lead-lag patterns in Figures 2 and B.3 is evident.

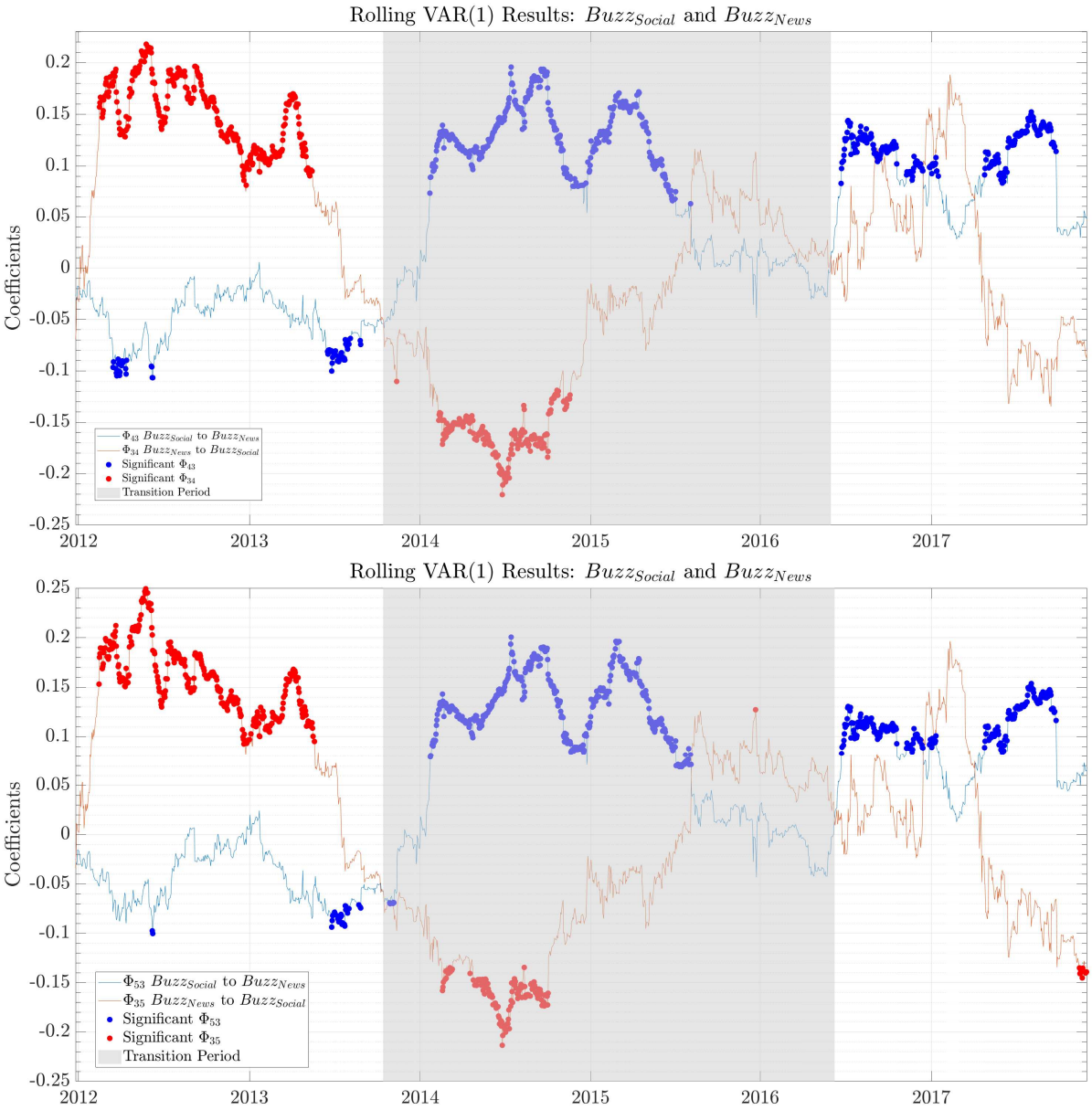


Figure B.2: Rolling Window VAR(1) Off-Diagonal Elements - daily *Buzz*. This plot depicts the inter-relationships between $Buzz_S$ and $Buzz_N$ series from 2011/01/01 to 2017/11/30. In contrast to Figure 1, where results are depicted for the bivariate system, $\mathbf{x}_t = (Buzz_S, Buzz_N)'$, in the current figure we present results from the four-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Buzz_S, Buzz_N)'$ in the top panel and a six-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Buzz_S, Sent_S, Buzz_N, Sent_N)'$ in the bottom panel. Sample contains 2,526 observations for each series, with the first 365 observations used as pre-estimation window. The shaded area indicates a transition period. The red line represents the leading effect from news media to social media, and the blue line indicates the leading effect from social media to news. Coefficients that are significant at the 90% level are shown with bold dots.

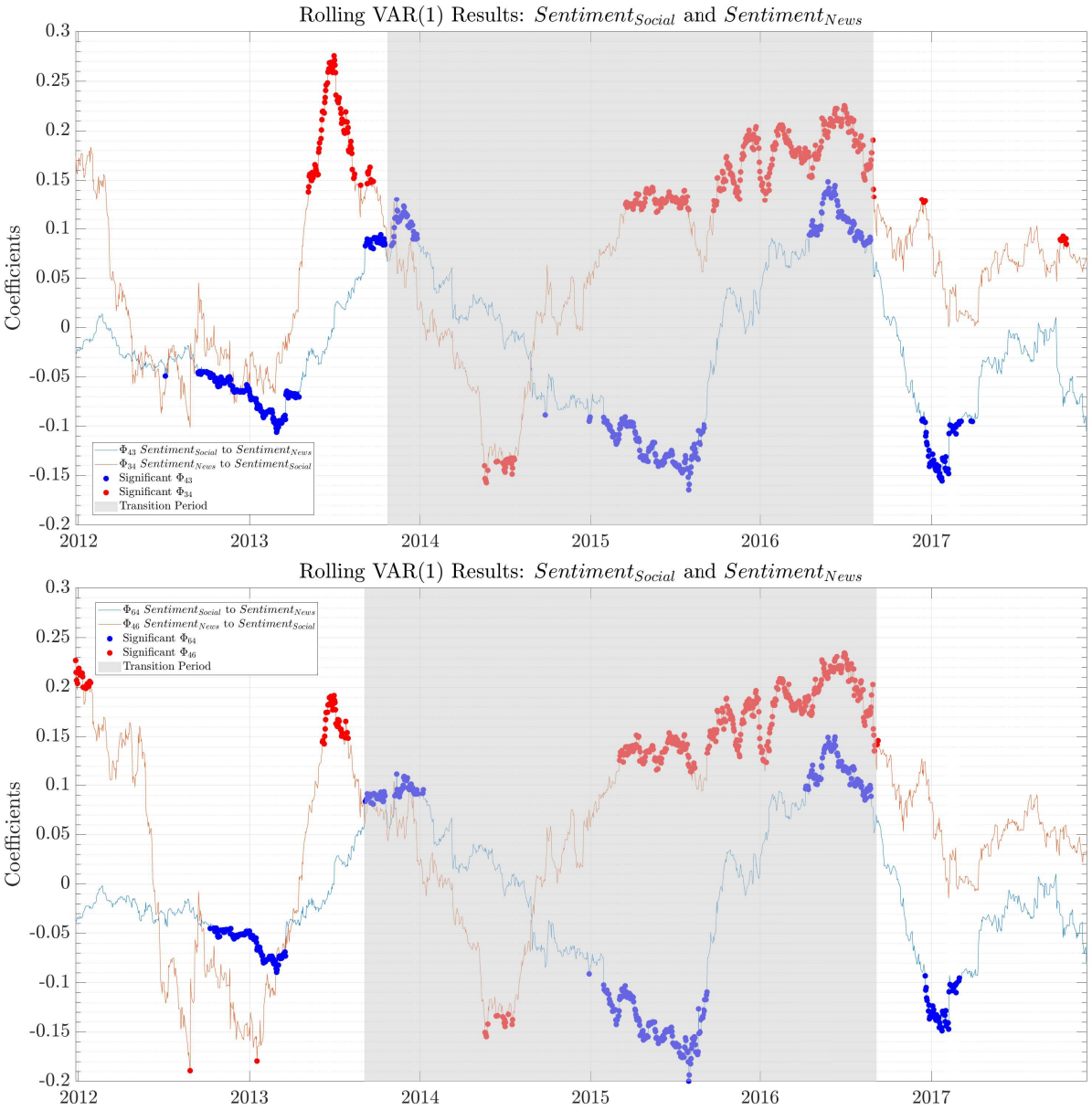


Figure B.3: Rolling Window VAR(1) Off-Diagonal Elements - daily *Sentiment*. This plot depicts the inter-relationships between $Sent_S$ and $Sent_N$ series from 2011/01/01 to 2017/11/30. In contrast to Figure 2, where results are depicted for the bivariate system, $\mathbf{x}_t = (Sent_S, Sent_N)'$, in the current figure we present results from the four-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Sent_S, Sent_N)'$ in the top panel and a six-variable VAR(1) system with $\mathbf{x}_t = (Return, VIX, Sent_S, Buzz_S, Sent_N, Buzz_N)'$ in the bottom panel. Sample contains 2,526 observations for each series, with the first 365 observations used as pre-estimation window. The shaded area indicates a transition period. The red line represents the leading effect from news media to social media, and the blue line indicates the leading effect from social media to news. Coefficients that are significant at the 90% level are shown with bold dots.

B.3 TRMI data and variables

Thomson Reuters MarketPsych Indices (TRMI) are derived from an unparalleled collection of prime news, global Internet news coverage, and a broad and reliable range of social media. The TRMI social media feed consists of both MarketPsych and Moreover social media content. Moreover Technologies' aggregated social media feed is derived from tens of thousands of social media sites and is incorporated into the TRMI from 2009 to the present. MarketPsych social media content was downloaded from public social media sites from 1998 to the present. After the social media posts or news articles are published in the TRMI content sources, a linguistic software abstracts the new content feed, parses and scores the content and attributes the score to global indices, companies, bonds, countries, commodities, currencies, and cryptocurrencies.

TRMI scores are composed of a combination of variables. The absolute values of all TRMI-contributing variables, for all asset constituents, over the past 24 hours are determined. These absolute values are then summed for all constituents. This sum is call the "Buzz".

Thomson Reuters MarketPsych computes 35 emotional scores which are divided into three types of sentiment indicators for a specific company or company group: 1) **Emotional** indicators including *Anger*, *Fear* and *Joy*; 2) **Fundamental** perceptions such as *Long vs Short*, *Earnings Forecast*, and *Interest Rate Forecast*; and 3) **Buzz** metric, a measure indicative of how much market-moving topics, such as *Litigation*, *Mergers*, and *Volatility* are being generated and discussed.

Thomson Reuters MarketPsych Indices (TRMI) analyse news and social media to convert the volume and variety of professional news and the internet into manageable information flows. The indicators are updated every minute for companies, sectors, regions, countries, commodities and energy topics, indices and currencies. TRMIs are based on relevant text collected over a window of content. If over that window there was no relevant text identified, then the correct value is "NA", not zero.[30] The indices are marked as ranging from either -1 to 1 (polarized indices) or 0 to 1 (unidirectional indices). TRMIs are evaluated on three different content sets: news, social media, and the combined content. History on all content dates back to the beginning of 1998. Only English-language text is used.

Collection of News media. Reuters news is present in the entire historical news dataset, as are a host of mainstream news sources collected by MarketPsych Data. During 2005, the archive began including Internet news content collected by Moreover Technologies. The Moreover content is restricted to those from top international and business news sources, top regional news sources, and leading industry sources.

Collection of Social media. The social media collection process is less diverse. It starts in 1998 with content collected by MarketPsych Data. Internet forums and finance-specific tweets compose this space. Starting in late 2008, Moreover Technologies social media content is included. Using popularity ranks measured by incoming links, this includes generally the top 30% of blogs, microblogs, and other social media content. Note that selected Moreover social media is included in the company groups social media dataset. The company groups data is composed of a subset of finance-specific Moreover content and the MarketPsych-based social media collection.

Tables B.5 and B.6 present descriptive statistics for the 35 sentiment indices based on social media and news respectively. We group **polarized ([-1,1])** and **unidirectional ([0,1])** emotional scores into Panels (A) and (B) respectively. The media activity measure, ***Buzz*** **($[0, \infty)$)**, is summarised in Panel

[30] An NA differs in meaning from true zero in that true zero represents the presence of text corresponding to positive and negative values that add up to zero. In other words, a zero value reflects that relevant text was found and its sentiment implications net to zero. In contrast, NA represents the absence of any relevant text and of any resultant measurement. Note that when the Buzz is zero, this means that no values were detected for any of the indices and thus all index values necessarily will be NA.

(C). All polarized sentiment scores are buzz-weighted, averaging any positive references net of negative references in the last 24 hours. Upon examination of the descriptive statistics, we observe the following facts:

- First, *Buzz*, a sheer media coverage volume metric for both social and news media, has a much larger absolute value than other emotional proxies (average *Buzz* value of 116,484.46 for social media and 202,401.31 for news, while other emotional scores contains mean value close to zero). Social media *Buzz* is highly positively skewed with the third moment equals to 1.37, and contains several large outliers. The kurtosis of 6.32 indicates a leptokurtic distribution (the last line in Table B.5). In contrast, news media buzz is more symmetric and contains less outliers than social media, with skewness equal to -0.01 and kurtosis 3.91 - slightly higher than 3 (the last line in Table B.6).
- Second, we observe fewer missing values among social emotional scores than among news in Panel (A) and (B), probably resulting from the fact that news reports require more stringent censorship procedures than social media. Peterson (2016, p.54) argues that "...Professional news sources include those with third-party editors and a journalistic responsibility to avoid slanderous or libelous commentary. Editors and fact-checkers ensure not only that news journalists uphold the brand's journalistic standards, but also that they do not commit libel or publish inaccurate information."
- Third, the [-1,1] polarized group scores from social media tend to be more extreme than the news. Buzz-weighted and normalised around zero mean, the polarized group emotional scores exhibit close mean and median values. However, the presence of large kurtosis values in the social media polarized group (Panel (A) of Table B.5) capture the large swings in emotional scores of social media posts. Similarly, although both social and news media unidirectional group indices suggest fat tail characteristics, extremely strong words are less frequent in news media than social media (Panel (B) of Table B.5 and Table B.6).
- Lastly, all of the TRMI indices are significantly autocorrelated with potential long memories. Our findings are based on Durbin-Watson (DW) test and Ljung-Box test with up to 5 lags (LB-5).

The availability of 35 emotional scores poses a dilemma: which emotional score is the most prominent one? In order to determine which emotional score(s) we should focus on, we report the **within group** pairwise contemporaneous correlations among all available sentiment indices in Figure B.5 on page 49 of the appendix. To aid interpretation and comparison of a large number of coefficients, we depict correlations in a schema ball instead of a large correlation table. Panels (a) and (b) depict associations among social and news indices, respectively. Yellow curves show positive correlations, and purple lines represent negative correlations. The thickness and brightness indicate the strength of correlation relationship, i.e. the thicker the curve, the closer the correlation coefficient is to ± 1. We find that, among both social and news based series, *sentiment* and *optimism* are strongly positive correlated with *marketRisk* - a measure defined by TRMI as "bubble-o-meter": the speculative extent relative to rationality. We also notice that *gloom* and *anger* embody the strongest negative correlations with *sentiment* and *optimism*. Therefore, we will pay closer attention to the following TRMI indices among the 35 available measures, namely: *buzz, sentiment, optimism, marketRisk, gloom*, and *anger*.

To measure the strength of dependence between social media and news based emotional scores, we employ Kendall rank correlation. Since emotional indices tend to sway from the normal distribution, the Pearson correlation is not appropriate. Using 500-day rolling window, Figure B.4 displays estimated correlation coefficients across time for the six indices mentioned above. Each line in the figure represents a correlation between an index based on social media and its news-based counterpart. The series are

positively correlated, indicating that social media and news-based scores are in concordance. The correlations, however, are far from perfect, validating our objective to contrast these two sources of investor sentiment. In addition, these concordance estimates exhibit strong heterogeneity across time, requiring analysis over several sub-samples.

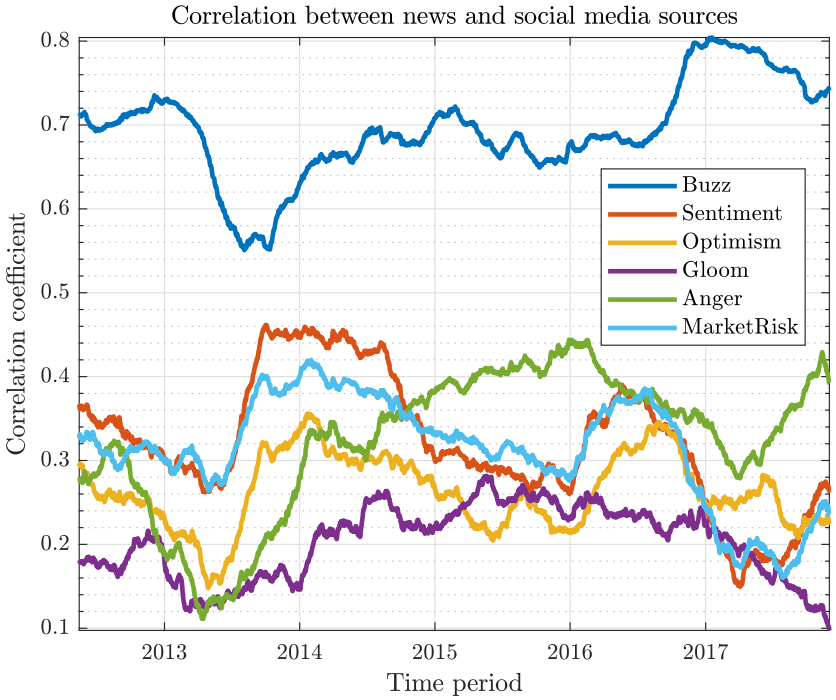


Figure B.4: Contemporaneous Correlation Dynamics between Key Social and News Indices. All six sentiment indices represent company group for the period from 2011/01/01 to 2017/11/30. Kendall correlation coefficients are calculated using rolling 500-day estimation window. For example, Buzz (blue line) depicts correlation dynamics between *buzz* from social media and *buzz* from news media. The correlation coefficients between social and news are positive for all six indices, however, they display time-varying heterogeneity over the sample period.

Based on these findings, we draw two conclusions that help us select the appropriate model specification. First, relatively low correlations suggest that social media and news do contain idiosyncratic components and that emotional scores based on these two types of media could be gainfully exploited either jointly or contrasted with each other in predictive regressions. Second, the time-varying relationship between social media and news-based indicators suggest that analysis should not be done over the entire sample period but rather with multiple sub-periods, e.g. a rolling window with a shortened span. In our quest to explore the lead-lag relationship between social media and news based sentiment, we further examine lagged cross-correlation (see graphs in Figure B.6 on page 50 of the appendix). Panel (a) displays the correlations between the previous day social media based indices and current day news indices, while panel (b) illustrates the correlation between the previous day news-based indices and the present day social indices. The findings are analogous to contemporaneous case: positively correlated social and news based series (although with lower magnitudes) and the time varying nature of lagged dependencies. Overall, Figure B.4 in conjunction with Figure B.6, indicate that the causal relationship between social and news media indices is dynamic, and causal modeling should be done in sub-samples rather than over the entire period.

We decide to focus on *Sentiment* and *Buzz* among all 35 indices as a result of both the above analysis and the Principal Component Analysis (PCA). We perform PCA separately on the polarized and unidirectional index groups (the list of all indices can be found in Tables B.5 or Table B.6). Since

Buzz metric is conceptually different from other emotional scores, we do not incorporate *Buzz* in the PCA analysis. To figure out how many principal components should be considered, we generate scree plots for social and news groups respectively in Figure B.7. Panels (a) and (b) depict the number of most influential components for the 18 polarized and 16 unidirectional social media group indices. The first principal component of polarized social sentiment indices explains 28.32% of total variances, and the second component explains an additional 10.76% of total variation (Panel (a) of B.7). The "elbow" appears at the second component, indicating that after the second principal component, incremental explanatory power of other components is greatly diminished. Likewise, the first principal component describes 22.19% of total group indices variances, and the second component constitutes an additional 10.71% of total variability. After the second primary component, the remaining components account for a very small incremental proportion of the variability and are probably unimportant (Panel (b) of B.7). Panel (c) and (d) illustrate the number of most influential components for TRMI news polarized and unidirectional emotional scores. For the polarized group ([-1,1]), the first component explains 29.51% of total variance, and the second component explains additional 12.70% (panel (c)). With respect to the unidirectional group [0,1], the first component accounts for 20.79% of total variance, and the second component facilitate to construe extra 11.77% of total variation (panel (d)). We observe that the "elbow" point also appear at the second component for news groups, indicating that after the second primary component, incremental explanatory power of other components decreases, thus they are less essential to our analysis.

Based on the findings above, we abstract the first two principal components and investigate each variable's contribution to these two principal components. To determine the most crucial variables among all TRMI indices available, we create biplots (see Figure B.8 on page 52) to assess the magnitude and sign of each variable's contribution to the first two principal components, and how each observation is represented in terms of those components. The axes in the biplot represent the principal components and the observed variables are represented as vectors. Figure B.8 in the appendix illustrates the results for both polarized (left panels) and unidirectional (right panels) sentiment scores based on social media (top panels) and news (bottom panels). Among the indices in the polarized groups, *Sentiment* and $emotionVsFact$ have the highest contribution to variation in both social media and news-based scores (Panel (a) and (c)). For unidirectional group, *violence* is the most prominent variable among the news-based scores (panel (d)), while for social media indices, there is no clear dominant component, instead a mix of *violence*, *stress*, *anger*, *gloom* and *joy* all playing incremental part in contributing to variation in unidirectional emotions from social media posts (panel (d)). We do not consider *violence* since we are focusing on the US market in this paper, although *violence* could be an important consideration for textual analysis research that investigates emerging markets or markets domiciled in geo-political and social unrest regions. Since involving multiple polarized emotional scores will hinder parsimony of our models, we decide to focus on *sentiment* and avoid entailing *emotionVsFacts* in our current framework.

Table B.5: Descriptive Statistics for TRMI MPTRXUS500 Company Groups based Social Media. Sample period 01/Jan/2011 - 30/Nov/2017; sentiment indices are grouped into polarized scores with [-1,1] range and scores that are unidirectionally bounded on [0,1]. *Buzz*, representing the volume of information flow, differs from other indices and is only bounded from below at 0. Data in *laborDispute* were too sparse over our sample period, but is included here for completeness. Results of Durbin-Watson and Ljung-Box (5 lags) tests indicates presence of autocorrelation in all indices.

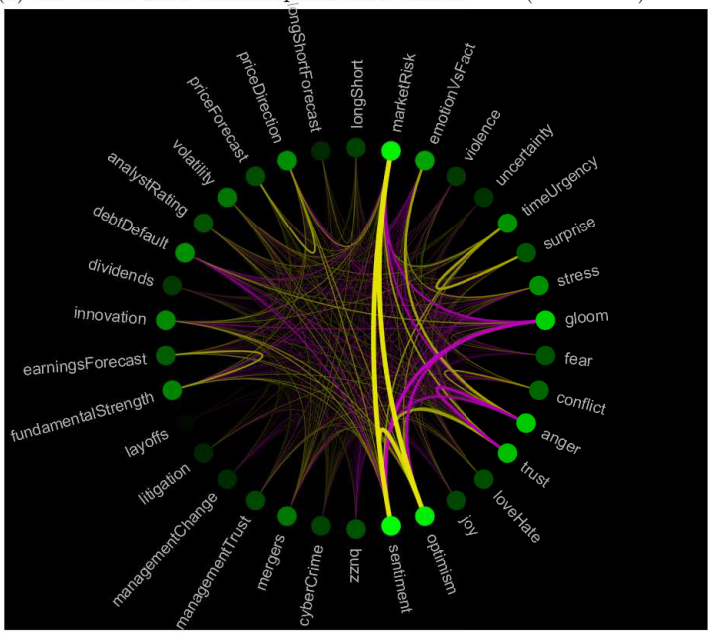
	Panel (A): Polarized Groups [-1,1]									
	Mean	**Std**	**Max**	**Min**	**Skew**	**Kurt**	**25th**	**Median**	**75th**	**IQR**
sentiment	-0.020	0.030	0.082	-0.127	-0.32	2.80	-0.040	-0.016	0.001	0.042
optimism	0.000	0.008	0.020	-0.034	-0.40	3.11	-0.005	0.001	0.005	0.010
loveHate	0.006	0.002	0.023	0.000	3.17	21.58	0.005	0.006	0.006	0.001
trust	-0.001	0.002	0.016	-0.021	-0.97	15.12	-0.003	-0.001	0.000	0.002
conflict	0.020	0.005	0.081	-0.002	2.70	21.92	0.017	0.020	0.023	0.005
timeUrgency	0.019	0.004	0.049	0.004	0.70	5.76	0.016	0.019	0.021	0.005
emotionVsFact	0.531	0.023	0.627	0.407	-0.20	4.54	0.518	0.532	0.546	0.029
marketRisk	-0.008	0.004	0.023	-0.027	-0.19	5.03	-0.011	-0.008	-0.005	0.005
longShort	0.004	0.004	0.090	-0.039	7.08	163.95	0.002	0.004	0.005	0.004
longShortForecast	0.001	0.001	0.003	-0.008	-1.87	24.97	0.000	0.001	0.001	0.001
priceDirection	0.003	0.002	0.014	-0.007	-0.04	4.33	0.002	0.003	0.004	0.003
priceForecast	0.001	0.000	0.003	-0.001	0.14	5.35	0.000	0.001	0.001	0.001
analystRating	0.001	0.001	0.008	-0.006	0.56	12.05	0.000	0.001	0.001	0.001
dividends	0.001	0.001	0.008	-0.004	2.12	25.00	0.001	0.001	0.001	0.001
earningsForecast	0.002	0.001	0.007	-0.003	0.86	6.03	0.001	0.002	0.002	0.001
fundamentalStrength	0.005	0.003	0.018	-0.004	0.86	4.73	0.004	0.005	0.007	0.003
managementChange	0.002	0.002	0.064	0.000	21.32	667.17	0.001	0.002	0.002	0.001
managementTrust	-0.001	0.002	0.016	-0.047	-7.58	114.09	-0.001	0.000	0.000	0.002
	Panel (B): Unidirectional Groups [0,1]									
	Mean	**Std**	**Max**	**Min**	**Skew**	**Kurt**	**25th**	**Median**	**75th**	**IQR**
anger	0.014	0.003	0.041	0.007	1.61	11.83	0.012	0.013	0.016	0.004
fear	0.005	0.001	0.010	0.003	0.98	6.86	0.005	0.005	0.005	0.001
joy	0.015	0.002	0.028	0.008	1.02	5.01	0.013	0.015	0.016	0.003
gloom	0.028	0.004	0.056	0.018	0.80	5.10	0.026	0.028	0.031	0.005
stress	0.056	0.004	0.099	0.044	1.35	15.43	0.054	0.056	0.058	0.004
surprise	0.008	0.001	0.026	0.005	2.23	21.96	0.007	0.008	0.009	0.002
uncertainty	0.023	0.003	0.035	0.012	-0.02	3.65	0.021	0.023	0.024	0.003
violence	0.029	0.005	0.063	0.021	1.90	8.72	0.026	0.028	0.031	0.005
volatility	0.026	0.003	0.055	0.019	1.47	10.56	0.024	0.026	0.028	0.004
debtDefault	0.004	0.001	0.018	0.002	2.07	15.73	0.003	0.004	0.005	0.001
innovation	0.003	0.001	0.011	0.001	1.02	6.48	0.002	0.003	0.003	0.001
laborDispute	-	-	-	-	-	-	-	-	-	-
layoffs	0.001	0.001	0.010	0.000	5.63	55.47	0.001	0.001	0.001	0.000
litigation	0.006	0.002	0.024	0.003	2.28	14.89	0.005	0.006	0.007	0.002
mergers	0.004	0.002	0.024	0.001	3.14	22.86	0.003	0.003	0.004	0.002
cyberCrime	0.001	0.001	0.015	0.000	5.53	47.44	0.000	0.001	0.001	0.001
	Panel (C): Buzz									
	Mean	**Std**	**Max**	**Min**	**Skew**	**Kurt**	**25th**	**Median**	**75th**	**IQR**
buzz	116,484.46	35,769.47	311,543.00	14,179.10	1.37	6.32	94,587.05	110,860.86	130,317.27	35,730.22

Table B.6: Descriptive Statistics for TRMI MPTRXUS500 Company Groups based News Media. Sample period 01/Jan/2011 - 30/Nov/2017; sentiment indices are grouped into polarized scores with [-1,1] range and scores that are unidirectionally bounded on [0,1]. *Buzz*, representing the volume of information flow, differs from other indices and is only bounded from below at 0. Data in *priceForecast, dividends, managementChange, laborDispute, layoffs* and *cyberCrime* were too sparse over our sample period, but is included here for completeness. Results of Durbin-Watson and Ljung-Box (5 lags) tests indicates presence of autocorrelation in all indices.

	Panel (A): Polarized Groups [-1,1]									
	Mean	**Std**	**Max**	**Min**	**Skew**	**Kurt**	**25th**	**Median**	**75th**	**IQR**
sentiment	-0.017	0.037	0.126	-0.173	-0.29	3.22	-0.042	-0.015	0.009	0.051
optimism	0.006	0.007	0.038	-0.037	-0.35	4.39	0.001	0.006	0.010	0.009
loveHate	0.005	0.001	0.013	0.000	0.69	7.18	0.004	0.005	0.005	0.001
trust	-0.001	0.002	0.006	-0.012	-0.86	5.49	-0.002	-0.001	0.000	0.002
conflict	0.032	0.006	0.056	0.017	0.87	4.07	0.028	0.031	0.035	0.007
timeUrgency	0.024	0.004	0.046	0.000	0.06	4.88	0.021	0.024	0.026	0.005
emotionVsFact	0.537	0.028	0.612	0.346	-0.68	4.40	0.521	0.539	0.557	0.036
marketRisk	-0.007	0.004	0.010	-0.031	-0.43	3.84	-0.010	-0.007	-0.004	0.005
longShort	0.002	0.003	0.014	-0.009	0.01	5.17	0.001	0.002	0.004	0.003
longShortForecast	0.000	0.001	0.003	-0.003	0.09	5.67	0.000	0.000	0.001	0.001
priceDirection	0.004	0.003	0.016	-0.012	-0.20	4.28	0.003	0.004	0.006	0.003
priceForecast	-	-	-	-	-	-	-	-	-	-
analystRating	0.001	0.001	0.007	-0.009	-2.16	21.26	0.000	0.001	0.001	0.001
dividends	-	-	-	-	-	-	-	-	-	-
earningsForecast	0.002	0.001	0.008	-0.004	0.60	4.56	0.001	0.002	0.003	0.002
fundamentalStrength	0.008	0.005	0.038	-0.005	1.48	7.35	0.005	0.007	0.010	0.005
managementChange	-	-	-	-	-	-	-	-	-	-
managementTrust	0.001	0.003	0.019	-0.017	-1.11	9.60	0.000	0.001	0.003	0.003
	Panel (B): Unidirectional Groups [0,1]									
	Mean	**Std**	**Max**	**Min**	**Skew**	**Kurt**	**25th**	**Median**	**75th**	**IQR**
anger	0.009	0.002	0.022	0.006	1.87	8.82	0.008	0.008	0.009	0.002
fear	0.007	0.001	0.014	0.004	1.19	6.56	0.006	0.006	0.007	0.001
joy	0.008	0.001	0.015	0.003	0.41	4.21	0.007	0.008	0.009	0.002
gloom	0.023	0.003	0.044	0.016	1.17	7.08	0.021	0.023	0.024	0.003
stress	0.056	0.005	0.078	0.042	0.58	4.09	0.053	0.055	0.059	0.006
surprise	0.007	0.001	0.020	0.004	2.15	17.83	0.006	0.006	0.007	0.001
uncertainty	0.019	0.002	0.030	0.012	0.43	3.52	0.017	0.019	0.021	0.003
violence	0.043	0.010	0.176	0.024	3.10	28.76	0.037	0.041	0.046	0.010
volatility	0.032	0.003	0.060	0.024	1.18	9.66	0.030	0.032	0.034	0.003
debtDefault	0.004	0.001	0.013	0.002	1.76	8.82	0.003	0.004	0.005	0.001
innovation	0.006	0.001	0.021	0.001	1.28	13.98	0.005	0.006	0.007	0.002
laborDispute	-	-	-	-	-	-	-	-	-	-
layoffs	-	-	-	-	-	-	-	-	-	-
litigation	0.011	0.003	0.038	0.005	1.60	9.33	0.009	0.010	0.013	0.004
mergers	0.005	0.002	0.022	0.001	1.68	9.49	0.004	0.005	0.006	0.002
cyberCrime	-	-	-	-	-	-	-	-	-	-
	Panel (C): Buzz									
	Mean	**Std**	**Max**	**Min**	**Skew**	**Kurt**	**25th**	**Median**	**75th**	**IQR**
buzz	202,401.31	47,847.27	387,635.55	1,468.90	-0.01	3.91	172,081.500	202,994.290	231,451.110	59,369.610

B.4 Correlation Schema-Balls

(a) MPTRXUS500 Contemporaneous Correlation (2011-2017) Social



(b) MPTRXUS500 Contemporaneous Correlation (2011-2017) News

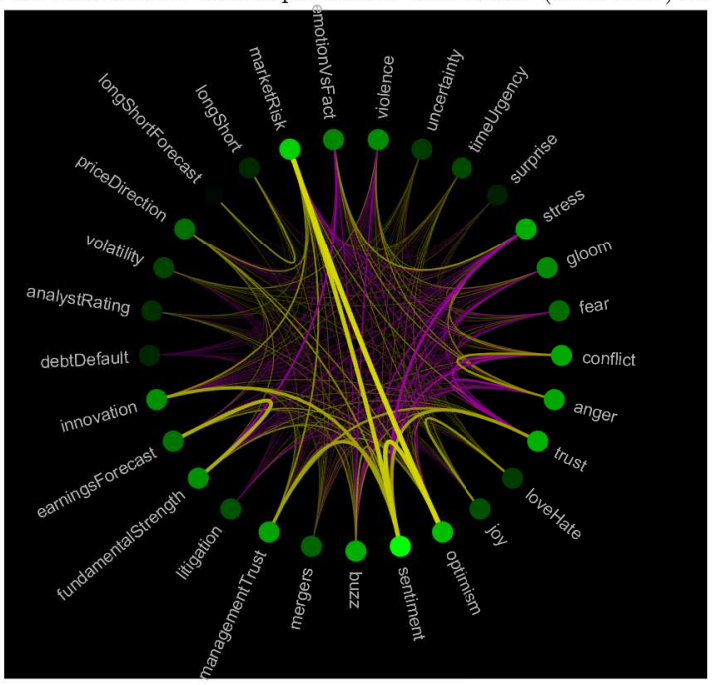
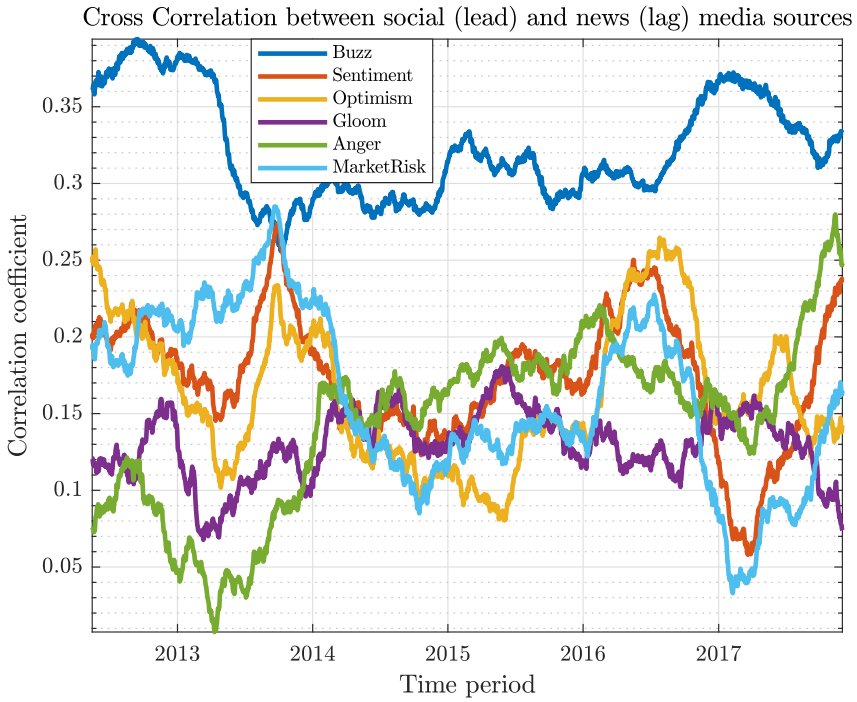
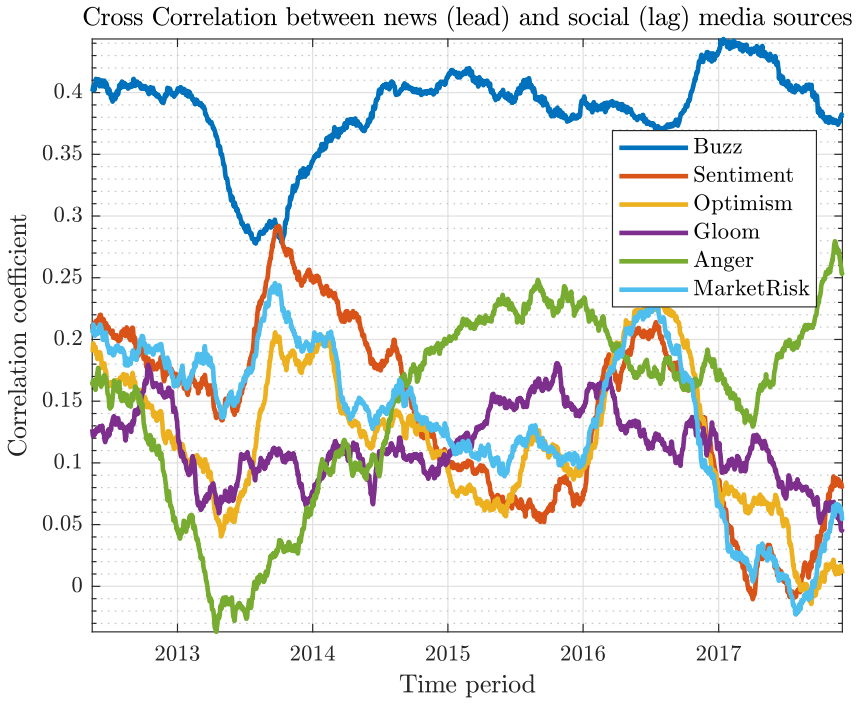


Figure B.5: Correlation coefficients between various emotional scores for the Company Group. The two panels are a visual representation of the pairwise contemporaneous correlations between all 35 scores for the company group (in place of 35-by-35 correlation matrices). Correlations for social media and news media based scores are highlighted in Panels (a) and (b) respectively. Yellow curves represent positive correlation coefficients, purple curves indicate negative correlations, the thickness and brightness of curves represent strength of correlation coefficients: the higher the absolute value of a correlation coefficient, the thicker and brighter is the curve that represents it. As indicated in Tables B.5 and B.6, there are more missing values among news-based scores. Concerned with the effect of data sparsity, we excluded a small number of emotional scores from our calculations. As a result, the number of variables in Panels (a) than (b) differ. Sample period: 01/Jan/2011 to 30/Nov/2017 at daily frequency.

B.5 One day lag cross correlations between social and news.



(a) *Social* leads *News* one day



(b) *News* leads *Social* one day

Figure B.6: One Day Lag Cross-Correlation between Key Social and News scores. Panel (a) shows Kendal correlation between key social and news scores for the Company Group based on daily data, i.e. the cross-correlation between $Social_t$ and $News_{t-1}$; Similarly, Panel (b) shows Kendal correlation between $News_t$ and $Social_{t-1}$. Both figures present similar patterns to Figure B.4 where the correlation between social and news based indices varies over time, suggesting an approach capable of capturing time-variability in the dynamics between social and news based emotional scores.

B.6 Principal Component Analysis

B.6.1 Scree Plots for Social and News Series

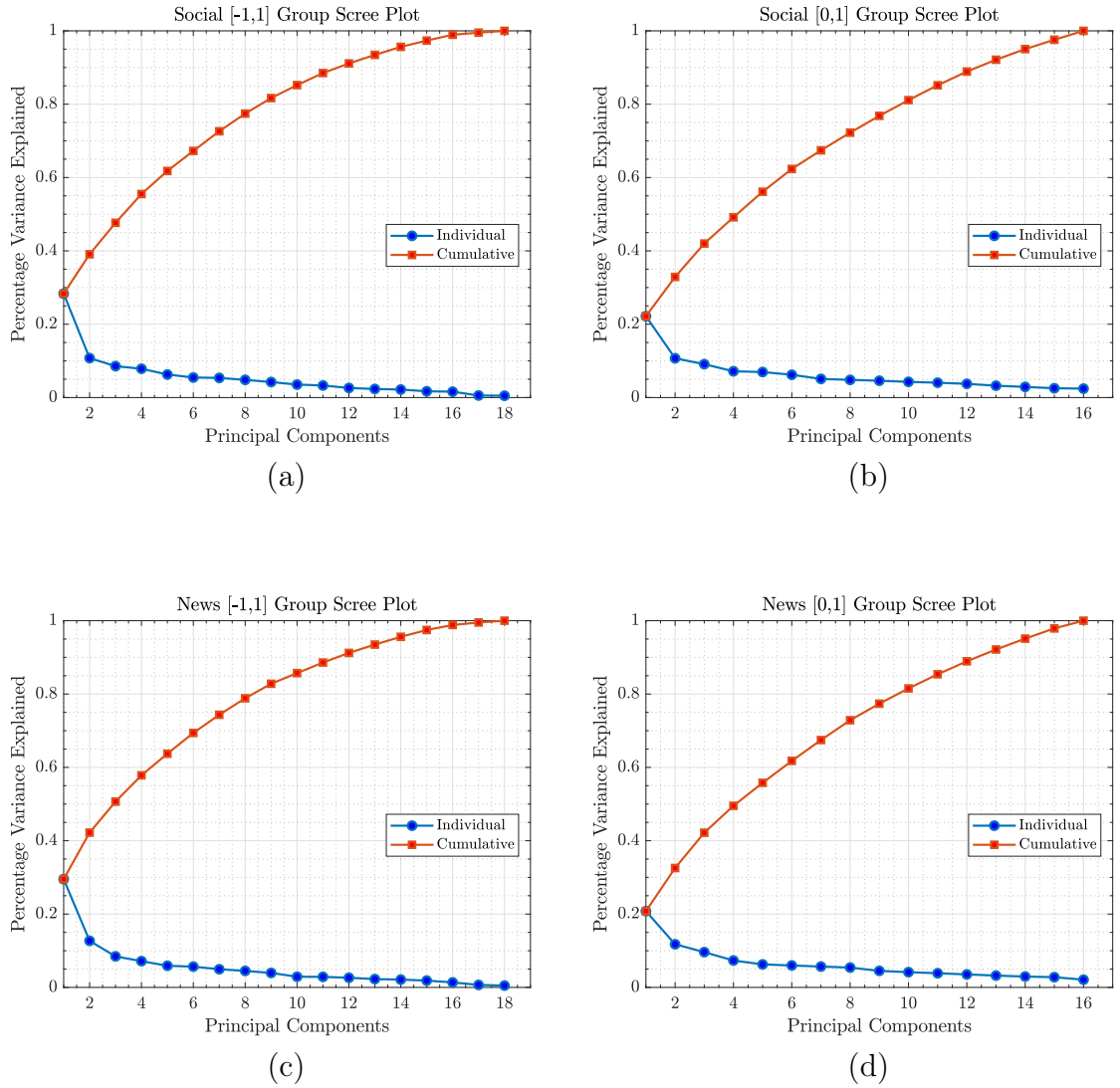


Figure B.7: Scree Plots from Principal Component Analysis of emotional scores for the Company Group. Panel (a) and (b) show individual (blue curve) as well as cumulative (red curve) contributions of each of the components considered based on PCA for the polarized group ([-1,1]) and unidirectional group ([0,1]) for **social** sentiment indices. For the polarized social sentiment indices (Panel (a)), the first component explains 28.32% of total variance, and the second component explains an additional 10.76% of total variation. For the unidirectional social sentiment indices (Panel (b)), the first component explains 22.19% of total variance, and the second component explains an additional 10.71% of total variation. After the second primary component, the remaining components account for a small incremental proportion of the variability and are probably unimportant. Panels (c) and (d) is constructed in a similar manner but based on **news** sentiment indices for the [-1,1] and [0,1] groups respectively. For the polarized news media group ([-1,1]), the first component explains 29.51% total variance, and the second component explains additional 12.70% (Panel (c)). With respect to the unidirectional news group [0,1], the first component accounts for 20.79% of total variance, and the second component facilitate to construe extra 11.77% of total variation (Panel (d)). Similar to social groups, after the second primary component, the remaining principal components account for a very small incremental fraction of the variability and are probably unimportant.

B.6.2 Biplots of the first two principal component coefficients

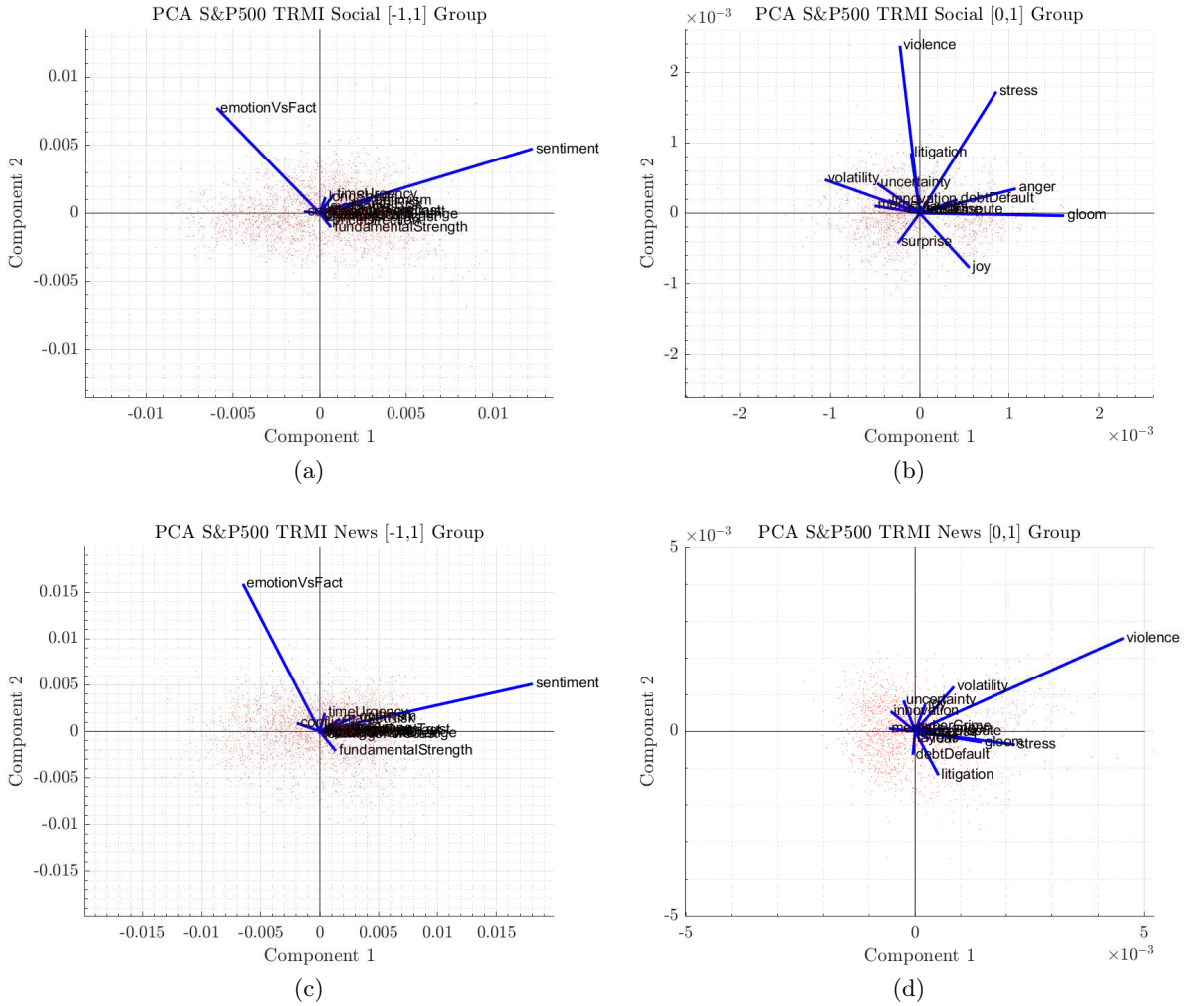


Figure B.8: Principal Component Analysis of the Sentiment Indices. Panel (a) is a biplot of the first two principal components for the [-1,1] sentiment score group in social sentiment indices; Panel (b) is a biplot of the first two principal components for the [0,1] sentiment score group in the social sentiment indices. Panels (c) and (d) are biplots constructed in a similar manner but using news sentiment data instead of social media. Panels (a) and (c) demonstrate that for both social and news media polarized groups ([-1,1]), *sentiment* and *emotionVsFacts* are the most crucial indices based on the variability they are able to explain in the data represented by the first two principal components. While Panel (d) indicates that *violence* is the most crucial emotional score in the news media [0,1] group, this conclusion is less obvious for the social media unidirectional group (Panel (b)). As *violence* is more relevant to research that focuses on emerging markets or markets that domicile in geopolitical unrest regions, we do not consider it in this paper. Since involving multiple polarized emotional scores will largely complicate the current research, we decide to focus on *sentiment* and avoid entailing *emotionVsFacts* in our models.

Hightlights:

- Using TRMI data, we discriminate social media from traditional news.
- We find that news media is dominant in earlier period between 2011 and 2013.
- Social media is becoming the dominant media source from 2016.
- Market return and volatility exert stronger impact on investor sentiment than the other way around.
- Link between volatility and sentiment is more persistent than the relationship between return and sentiment.