Sensitivity to Sentiment: News vs Social Media[*]

Baoqing Gan[†], Vitali Alexeev, Ron Bird, and Danny Yeung

Finance Discipline Group, UTS Business School, Sydney, Australia

April 30, 2019

Abstract

Applying sentiment indicators generated by an algorithm that extracts texts from major social and news media outlets, this paper investigates how social and news media interact with each other from 2011 to 2017, and analyses the overall market sensitivity to media sentiment under different information environments. Our rolling-window VAR models find that quantity and sentiment of news media significantly lead social media before 2014, while social media plays a leading role to news after 2016. Further analysis at each sub-sampling period corroborates these transition patterns and provides new evidence that the effects from market return and volatility to media sentiment are stronger than the influences from sentiment to market. We also observe that the revision of mispricing induced by media sentiment have expedited after 2016, and that impacts from media sentiment on volatility are more persistent than on return. Overall, this paper directly contrasts social and news media effects and contributes to literature on investor sentiment and noise trader risks that uses synthesized media textual analysis data.

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

> *"Public sentiment is everything. With public sentiment, nothing can fail. Without it, nothing can succeed."*
>
> —*Abraham Lincoln*

[*] 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.

[†] PhD student and casual academic at UTS Business School, Email: *baoqing.gan@uts.edu.au*; Phone: +61 2 9514 7787. This research constitutes a part of the PhD thesis. I am grateful to the continuing support, guidance and encouragement from my supervisory panel: Vitali Alexeev *vitali.alexeev@uts.edu.au*, Ron Bird *ron.bird@uts.edu.au*, and Danny Yeung *danny.yeung@uts.edu.au*.

1 Introduction

The financial sentiment literature has shown that macroeconomic announcements, major geopolitical events, and corporate announcements change investors' sentiments and often influence stock prices. Traditionally, investors receive this information through mainstream financial news reports, official announcements, corporate conference calls, and analysts research reports. Nowadays, social media outlets such as StockTwits and internet message boards play a more prominent role in the information dissemination process, delivering greater quantities of company related information to the market at faster speeds. Social media, however, has been known to create attention-grabbing hot topic that may sway investors' beliefs about company's future outlook, thus forming investor sentiments that ultimately impact stock prices. 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). As early as 1896, Le Bon (Le Bon, 1896) pointed out that when people are in certain groups, they will behave quite differently from when they are alone. Group sentiments are contagious. Individual's behaviour varies in accordance with their social contexts. Similarly, Fehr and Tyran (2005) found that a small amount of individual irrationality may lead to large deviations from the aggregate predictions of rational models under certain circumstances. News media plays an important role as the storyteller and information-transmitter for social interactions, which ultimately influence the stock market dynamics.

Recently, 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] In early 2013, Bloomberg announced that it would add Twitter accounts to its financial information terminals - a "must-be" tool used by traders on Wall Street.[2] 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 Syrian 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; the S&P 500 temporarily lost 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.

The US Securities and Exchange Commission (SEC) has been keeping up with the trend: after issuing a guidance in 2008 admitting that corporate websites can serve as an effective means for

[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>.

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". This investigation report stems from the CEO of Netflix posting on his Facebook account that Netflix's monthly online viewing had exceeded one billion hours for the first time, without disclosing this information through Form 8-K or other press releases. Accordingly, the stock price of Netflix increased from $70.45 at the time of the Facebook post to $81.72 at the close of the following trading day.[4] Continuing to warm social media up, 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" in June 2015 (Bartov et al., 2018).

The changes in financial industry brought out by new technology motivate us to address two main research questions concerning propagation and dynamics of investor sentiment. Firstly, we aim to find out: how social and news media interact with each other over time? Specifically, we develop bivariate rolling-window vector autoregressive (VAR) models to analyse the dynamic in the lead-lag relationship between social and news media from 2011 to 2017. Using the sheer volume of social/news media activity, our first model investigates whether increases in volume of the news press coverage, $Buzz_N$, lead to higher social media postings, $Buzz_S$. In addition, accounting for the tonality contained in textual data from social/news media posts, the second model explores whether the net positive and negative sentiment from news media, $Sent_N$, drives subsequent net emotions in social media, $Sent_S$. We find that between 2011 and 2014 both the coverage and the sentiment from news media stimulate their social media counterparts. Beginning from 2016, however, changes in volume and sentiment from social media impact the news media more strongly. The period from 2014 to 2016 is identified as transition period with mutual causation in social and news media activities where the dominant role of one information channel over the other is less discernible. This transition period is closely aligned with SEC rulings on legitimacy of social media platforms as companies' official information channels.

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 2014 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 reveal that the speed of reactions for both return and

[4] The US Securities and Exchange Commission 2013, *SEC Says Social Media OK for Company Announcements if Investors Are Alerted*, Press Release, accessed 12 August 2018, <https://www.sec.gov/news/press-release/2013-2013-51htm>

sentiment have accelerated after 2016 compared to the period before 2014. Return responses to a shock from social media sentiment almost doubled after the transition period (from 0.03 to 0.07), while the return responses to one standard deviation change in news-based sentiment shrunk to about half of 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-2014 and post-2016 periods displays higher sensitivity to social media sentiment than to news-based sentiment. Social media shocks to volatility process are more persistent compared with returns with the reversion process taking in excess of 20 days. We conclude that the media does not follow market variations 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 and observe propagation of social media in later periods resulting from recent developments in information and telecommunication technology as well as acceptance of the new technology by regulatory authorities. Our results suggest that researchers in this topic should and must consider the time-varying nature of the relationship between social and news media. To the best of our knowledge, there is no other research that highlights such differences and details sentiment effects from different media sources on stock market. 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 synthesize 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 follow: 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 sentiments and stock market return/volatility. We conclude in Section 6 and propose directions for future research.

2 Literature Review

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, including 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.[5]

2.1 Investor Sentiment and Stock Market

Early research on investor sentiment and stock market movements are generally based on sentiment created from market fundamental variables. For example, the influential Baker & Wurgler sentiment index (Baker and Wurgler, 2006) abstracts six prominent variables using Principal Component Analysis (PCA) approach. The sentiment indicator from the American Association of Individual Investors (AAII) is compiled from direct survey results.[6]

Using these sentiment proxies, a variety of studies test behavioural finance theories such as security market under-/overreactions at both aggregate market and individual stock levels. Such "behaviour augmented" models usually consider various investor heuristic bias, for example, overconfidence and self-attribution bias by Daniel et al. (1998), conservatism and representativeness by Barberis et al. (1998), and confirmation bias by 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. 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 variations 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

[5] 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.

[6] For review of other fundamental variable based sentiment measures, refer to Baker and Wurgler (2007).

(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 Stambaugh et al. (2012) find no predictability of 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, such as Baker & Wurgler and AAII, that are 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.

2.2 Investor Sentiment Based on Textual Analysis

In recent decade, a growing body of literature are assisted by the advancement of textual analysis and machine learning techniques. Following the pioneering research from Tetlock (2007) and Tetlock et al. (2008), which collect and assign emotion scores of text messages from a specific column in the *Wall Street Journal*, finance academics are paying more attention to the relationship between stock market and information quantity, as well as sentiment conveyed within textual data. Based on the term-weighting schemes from Loughran and McDonald (2011b), four main information sources are examined by researches: **corporate filings**, **professional financial news presses**, **internet message boards**, and **social media platforms** such as Twitter and StockTwits.[7]

Empirical research relying on scanning and scoring texts from filed documents and press releases is abundant and still expanding. Using 10-K words tonality, Loughran and McDonald (2011a) and Jegadeesh and Wu (2013) investigate the filing period drift. Applying the computational linguistics methods to classify texts from *Wall Street Journal* columns, Antweiler and Frank (2006) explores the impact of negative attitudes in news press on corporate events. Engelberg (2008) differentiates two types of information: the hard quantitative information and the soft qualitative information, and tests their effects on post earnings announcement drifts using corporate earnings announcement news from Dow Jones News Service (DJNS) on *Factiva*. Fang and Peress (2009) looks into four major newspapers: the *New York Times*, *USA Today*, *Wall Street Journal*, and *Washington Post* to study the relationship between media coverage and market reactions. They find evidence supporting the "thinly covered stock premium" hypothesis of Merton (1987). Engelberg et al. (2012) explores how short-sellers gain their information advantage by examining texts from Dow Jones News Service and the *Wall Street Journal*. Creating a specific positive and negative emotion index from two columns of financial news from the *New York Times*, Garcia (2013) looks into the linkages between media sentiment and the market from 1905 to 2005 during major recessions. Interestingly, they show that, controlling for other well-known time-series patterns, the predictability of stock returns using news' content is concentrated in recessions.

[7] 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).

Most of the empirical work focuses on either the volume (e.g., coverage) or the sentiment (positive/negative emotions) 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.

In the last few years empirical research has shifted its focus to analysing effects of social media on the stock market. Concentrating on sentiment from internet message board, Wysocki (1998) demonstrates that the quantities of *Yahoo!Finance* posts have predictive power for the next day trading volumes. Antweiler and Frank (2004) analyses messages from *Yahoo!Finance* and *RatingBull*, and detects its interrelations with stock market return, volatility, volume and bid-ask spread. The authors find that talks on the web forums do not have price predictability but do sway the volatility. Accounting for the slang and the ambiguity of language on the web, Das and Chen (2007) improves the sentiment extraction process by developing their own algorithm that scrapes message board postings. Empirical tests applying this new approach on 24 Morgan Stanley High-Tech firms indicates that sentiment does contain certain explanatory power to the aggregate stock price level and a diminished power to the price changes. This result, however, suffers from small sample bias since it only considers a two-month period from July to August 2001. Focusing on peer opinions and non-professional investors' communications, Chen et al. (2014) checks commentaries conveyed in the stock discussion forum, *SeekingAlpha*, and finds that the sentiment derived from these postings displays predictability for stock return and earnings, and that the high frequency of negative words is associated with more prominent predictability.

Based on sentiment extracted from firm-level Tweeter posts of S&P 100 companies, Sprenger et al. (2014b) analyses the inter-reactions between tweets' features (bullishness, posting volumes, and agreement) and stock market features (return, volume, and volatility). Using Fama-MacBeth regressions, they find that the feedback effect from stock market to social media variables prevails. Similarly, Ranco et al. (2015) investigates the relation between Dow Jones Industry Average (DJIA) firm tweets and stock performances. They find that the cumulative returns around earnings announcements are dependent on Twitter sentiment during high posting periods. Following this line of research, textual sentiment generated from other social media sources started to emerge. Da et al. (2011) uses Google Search Volume (GSV) Index of Russell 3000 as a superior measure of investor attention, and discovers a for-nightly upward stock price drift after surging GSV index, which helps explain the IPO first day excess return anomaly. Siganos et al. (2014) uses daily sentiment index from Facebook's *Gross National Happiness* to show that this index is positively, but unfortunately only contemporaneously, correlated with stock return, and that negative values of this index coincides with spikes in trading volume and volatility.

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. 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. Das and Chen (2007) examines 24 high-tech companies in the two-months period from July to August 2011. 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) constructs a proprietary algorithm that abstract textual sentiment from 750,000 StockTwits at intra-day level between September 2014 and April 2015 and finds that the first half-hour sentiment changes manifest market return predictability to the last half-hour.

2.3 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 contains predictability up to a quarter. They also find asymmetric reversal process for positive and negative news sentiment.

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. To do this, we aggregate the daily TRMI social media and news sentiment scores (denoted as $Sent_S$ and $Sent_N$ respectively)[9] into monthly frequency and report the correlations between TRMI and the BW sentiment indices in Table 1.[10] The results in Table 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

[8] A detailed summary of this dataset and description of our sample is presented in Section 3.

[9] A list of variable and acronym can be found in Table A.2 on page 36 of the Appendix.

[10] 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/>.

within sentiment scores (news vs social) and between sentiment and market variables (sentiment vs returns and/or volatility)..

Table 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 BW index. BW sentiment data is obtained from personal website of Jeffrey Wurgler on 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

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 sentiments.

Our paper is complimentary to Sun et al. (2016), Nooijen and Broda (2016), and Jiao et al. (2018) 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 S&P 500 Index. They substantiate that the first half-hour sentiment changes from TRMI are helpful to forecast the last two hours' stock index return, which is different from within day momentum effect. They point out that this predictability is able 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 et al. (2018) 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 the coming month.

This paper distinguishes itself by contributing to the literature from two main points. Firstly, similar to Jiao et al. (2018), we contrast 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. In doing so, we address the important question: did the media landscape change 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 MarketPych 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 S&P 500 company group. Our S&P 500 stock market data is obtained from Datastream. 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.

TRMI offers 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

[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 subjests, 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.

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.

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 for the *MPTRXUS500* company group data from 2011 to 2017. *MPTRXUS500* index aggregates sentiment and tone of the largest 500 companies in the US, and aims at capturing the S&P 500 index sentiment. The data are updated each day at 3:30pm US Eastern time, including weekends and other non-trading days.[12] Tables 2 and 3 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*** $([\mathbf{0}, \infty))$, is summarised in Panel (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 2). 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 3). 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. 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 2) 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 2 and Table 3). Lastly, all of the TRMI indices are significantly autocorrelated with potential long memories.[13]

[12] Further details on the TRMI data can be found in the Marketpsych white paper by Peterson (2013).

[13] 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.

Table 2: 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 3: 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

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 A.1 on page 40 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* embodies 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 1 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.

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 A.2 in page 41 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 1 in conjunction with Figure A.2, 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 from both the

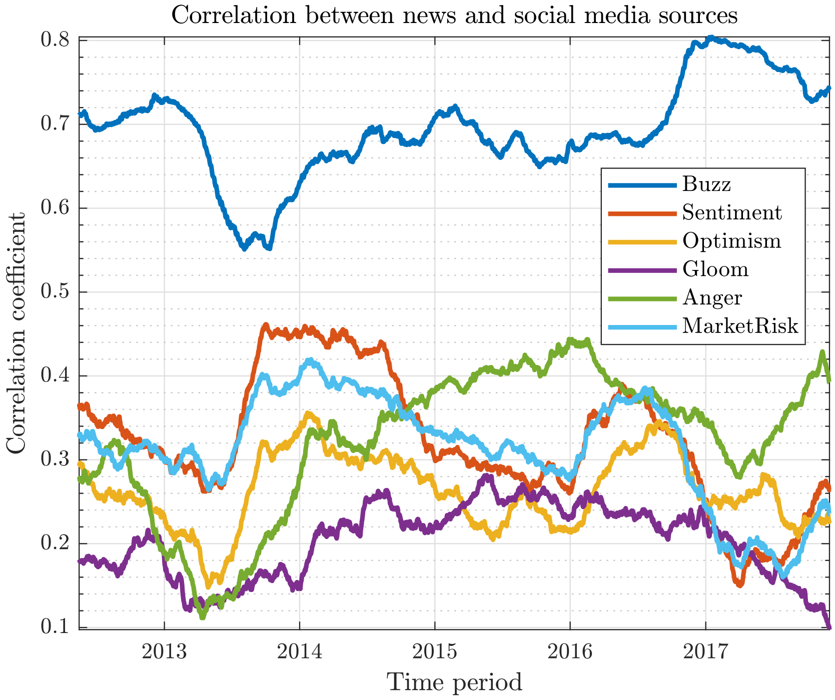


Figure 1: Contemporaneous Correlation Dynamics between Key Social and News Indices. All six sentiment indices represent S&P 500 company group for the period from 2011/01/01 to 2017/11/30. Kendall correlation coefficients are calculated using rolling 500-day estimation window. For example, Buzz (blue line) depicts correlation dynamics between *buzz* from social media and *buzz* from news media. The correlations coefficients between social and news are positive for all six indices, however, they display time-varying heterogeneity over the sample period.

above analysis and by performing 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 2 or Table 3). 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 A.3 in the appendix. 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 A.3). 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 A.3). 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 decrease and 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 A.4 on page 43 in the appendix) 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 A.4 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.

3.2 Stock Market Data

The sample period for the stock market data is consistent with the availability of our TRMI data, e.g from 01/Jan/2011 to 30/Nov/2017 sampled daily. Fortunately, this period avoids the turmoil of the global financial crisis (GFC) episodes from 2008 to 2010. At the same time, this sample period covers a phase of rapid development of the social media, thus permits us to compare and contrast 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 4:[14]

Table 4: Descriptive Statistics for the S&P 500 Index over the period 2011/01/01-2017/11/30. Returns are calculated as $r_t = log(\frac{P_t}{P_{t-1}})$, where P_t is the daily close price for the S&P 500 index obtained from Datastream. Reported figures are annualized by multiplying the daily return values by 252. VIX data is acquired from WRDS CBOE S&P 500 volatility index futures closed prices. The unreported Durbin-Watson test and Ljung-Box 5 lags test for all indices show presence of autocorrelation for both return and VIX series.

	Mean	Std	Max	Min	Skew	Kurt	25th	Median	75th	IQR
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

[14] A full list of all data sources is available in Table A.2 in the appendix.

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.

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 A.5 and A.7 (pages 44 and 45). 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.[15] To deal with weekly effects in *Buzz* and *Sentiment* series, we regress *Sentiment* and winsorized *Buzz* on day-of-the-week dummy variables, retaining fitted residuals as our seasonally adjusted data. Figure A.6 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.8 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 A.6 and A.8 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.[16] VAR provides a simple framework systematically capturing rich dynamics in multiple time-series. We rely on two derivative frameworks: a rolling-window VAR method and structural VAR (SVAR) model 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

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

[16] 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”.

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:[17]

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.[18] 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.);

$\mathbf{x}_t$ 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).[19] BIC generally penalizes free parameters more strongly than AIC, allowing for more parsimonious models.

4 News vs Social Media: Dominating Causality Pattern

We examine the serial dynamic relations between $Buzz_S$ and $Buzz_N$ by estimating a VAR model using S&P 500 TRMI company group data. We choose S&P 500 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

[17] A full list of variables, the notations and definitions of them used in this study is available in Table A.2.

[18] $\{\epsilon_t\}$ is also called impulse, or innovations (Tsay, 2005).

[19] For notation and definition details, refer to Table A.2 in the appendix.

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 2 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 with bold points.[20] 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.

From Figure 2, we observe that the blue and red coefficients crossed in October 2013. Prior to this "transition" point, the magnitude of red line (Φ_{21}) is above blue line (Φ_{12}), with more numbers of Φ_{21} coefficients being significant than the Φ_{12} coefficients. For example, in Table 5 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 is significant at 1% level. By contrast, ϕ_{21}, the impact from $Buzz_S$ to $Buzz_N$, is -0.0329 and is 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 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_{12} (the blue line, social to news) trends further upward, remaining significant, while Φ_{21} (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 5, 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 is not statistically significant. This result confirms the

[20] Based on our analysis, a VAR model with 7 lags is optimal according to BIC criterion. Detailed AIC and BIC results for this system is available in Appendix Table A.3 Panel A, page 37. However, we report VAR(1) as it is a parsimonious form of VAR(7) based on the model specification test shown in Table A.4, page 38. According to Table A.4, most of the inter-mediate 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 Information Criteria might be determined by 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.

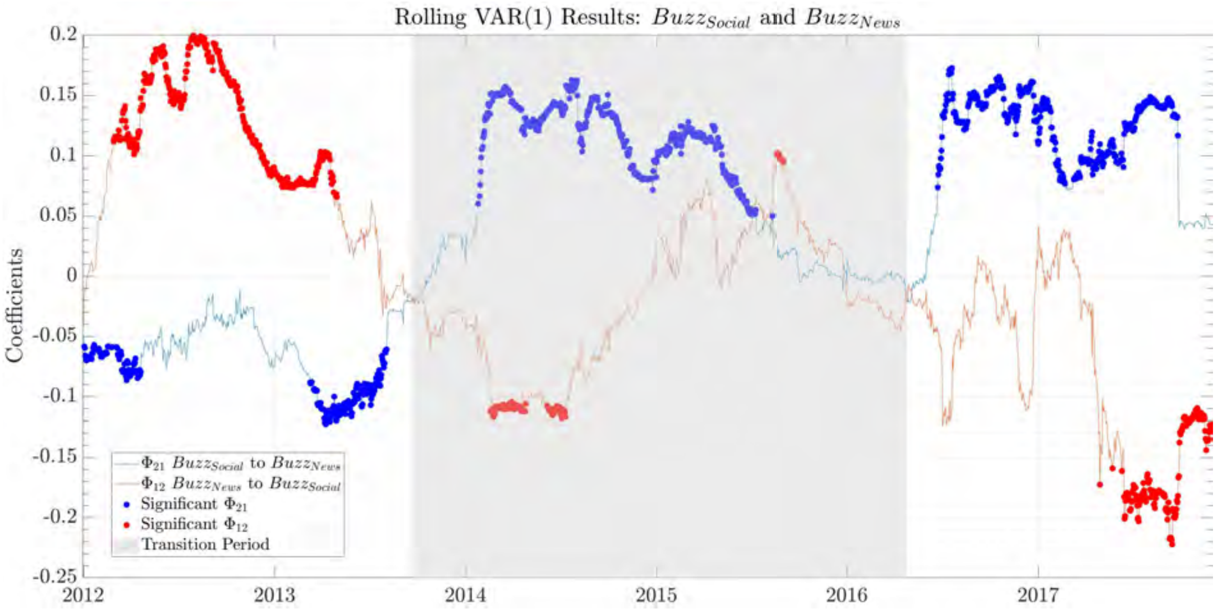


Figure 2: Rolling Window VAR(1) Off-Diagonal Elements - daily *Buzz*. This plot depicts the inter-relationships between $Buzz_S$ and $Buzz_N$ series from 2011/01/01 to 2017/11/30. Sample contains 2,526 observations for each series, with the first 365 observations used as pre-estimation window. The shaded area indicates a transition period. The red line represents the leading effect from news media to social media, Φ_{12} in equation system (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.

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 that now plays a predominant role in the flow of information.

Table 5: Before vs After Transition Period VAR Slope Coefficients: Social vs News. Panel A and B reports 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 frequency 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 The left side result is one representative regression performed in the pre-transition period, and the right side result is one typical regression conducted in the post-transition period.

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

	Pre-transition Period				Post-transition Period				
	Value	SE	t-Stat	p-Value		Value	SE	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				
	Value	SE	t-Stat	p-Value		Value	SE	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)[21]. In Figure 3, 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 5, 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 of Figure 3, we find that in spite of some fluctuations in the transition period when news and social mutually influence each other, we can see that 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 5 indicates that one of the "Post-transition Period" VAR has social to news effect (ϕ_{21}) equal to -0.1589 and is significant at 1% level. But news sentiment influences social (ϕ_{12}) insignificantly (p-value of 0.73) at a lower level (0.0166). This result is consistent with the pattern we identified in Figure 2. In both figures, news media impacts are leading social media effects before the transition period, however, after the transition period this pattern is reversed.

[21] Table A.4 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 3

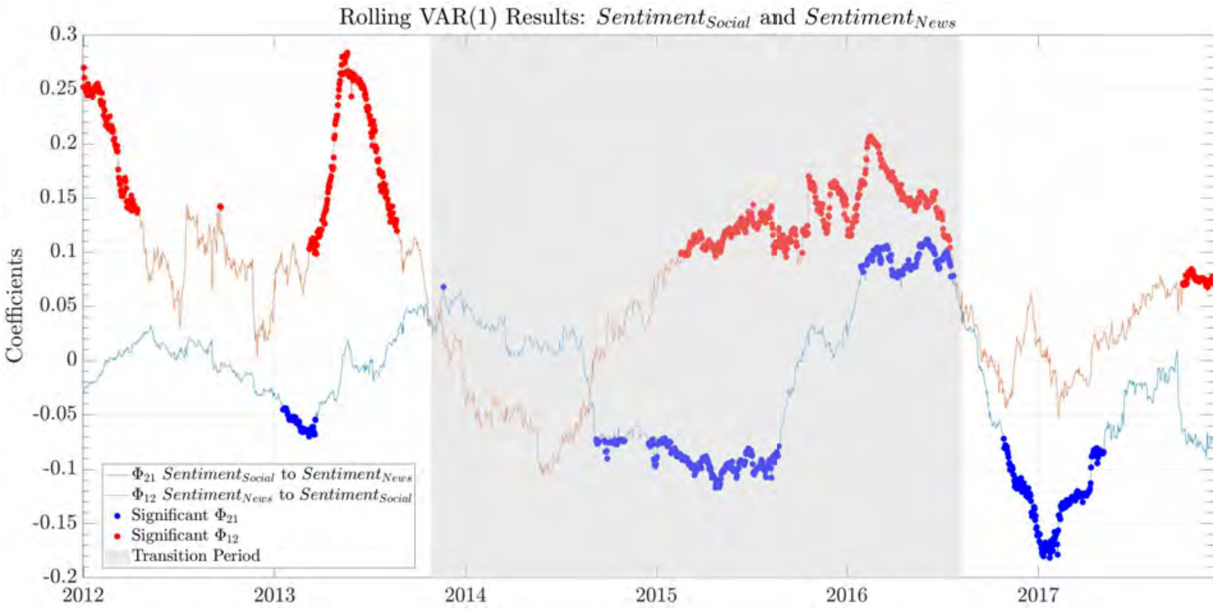


Figure 3: Rolling Window VAR(1) Off-Diagonal Elements - daily *Sentiment*. This plot depicts the inter-relationships between $Sent$ and $Sent_N$ series from 2011/01/01 to 2017/11/30. Sample contains 2,526 observations for each series, with the first 365 observations used as pre-estimation window. The 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 structural 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 reverse impacts from return to media sentiments. 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.[22]

Panels A and B in Table 6 summarise the results for VAR systems in (3) and (4) respectively over pre-transition 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-transition Period and Post-transition Period in Panel A and B are all positive and significant at 5% level: 0.0995 in the Pre-transition Period model, and 0.1929 in the Post-transition Period model for the social media sentiment VAR system in Panel A; 0.1060 in the Pre-transition Period model and 0.2171 in the Post-transition Period model for the news sentiment regression in Panel B. Theses results indicate that return has positive and significant impacts to both social and news sentiments. In contrast, initial sensitivities of return to sentiment, the ϕ_{21} coefficients in panel A and B, are insignificant for all four estimations. This results corroborates findings in prior literature that sentiment is more sensitive to return shocks than return is to sentiment shocks (e.g. Sprenger et al. (2014a)). However, these values could not canvas for the dynamic process of responses for the dependent variable from shocks in the leading variables. To better observe the results and contrast social media effects with news, we generate Impulse Response Functions (IRFs) for the leading 20 working days (one month) of this group of estimations in Figure 4.

The left side plots in Figure 4 represent IRFs that capture return responses to social or news media sentiment shocks. Panel (a) and (c) indicate responses of return to **social** sentiment shocks in the Pre-transition Period and Post-transition Period respectively, whereas panel (e) and (g) represent responses of return to **news** sentiment shocks in these 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 the deviations revert back to zero gradually at different speeds. This finding is consistent with the overreaction explanation, which proposes that sudden surges in investor sentiment lead to temporarily spikes in stock prices that will retreat shortly.

A comparison of panel (a) with panel (c), reveals two interesting findings. First, the influence to return from social media sentiment increased after the transition period. In particular, the magnitude of IRFs goes up from 0.03 before 2014 to 0.07 after 2016 - the sensitivity almost 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 arrives back to its original level in about 3 weeks (15 working days), while in the post-transition period, return bounces back to normal 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 Period level of 0.030 (panel (e)) to the Post-transition Period of 0.016 (panel (g)) - approximately halved in value. However, similar to the social media effects, the speed of

[22] As is shown in Table A.3 in the Appendix, VAR(5) is optimal for these two systems according to BIC. However, we report VAR(1) results in Table 6 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 6: Before vs After Transition Period VAR Slope Coefficients: Sentiment Vs Market. Panel A to Panel D reports 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 stock return to social media sentiment, while ϕ_{21} shows the impacts from social media sentiment to stock market 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 panel 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 measures as the squared term of the seasonality adjusted and non-trading day averaged *Sentiment* series. ϕ_{11} and ϕ_{22} are the autocorrelation for $Sent^2$ and VIX in panle C and D. For each panels from A to D, the left side result is one representative regression performed in the pre-transition period, and the right side result is one typical regression conducted in the post-transition period.

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

	Pre-transition Period					Post-transition Period			
	Value	SE	t-Stat	p-Value		Value	SE	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*

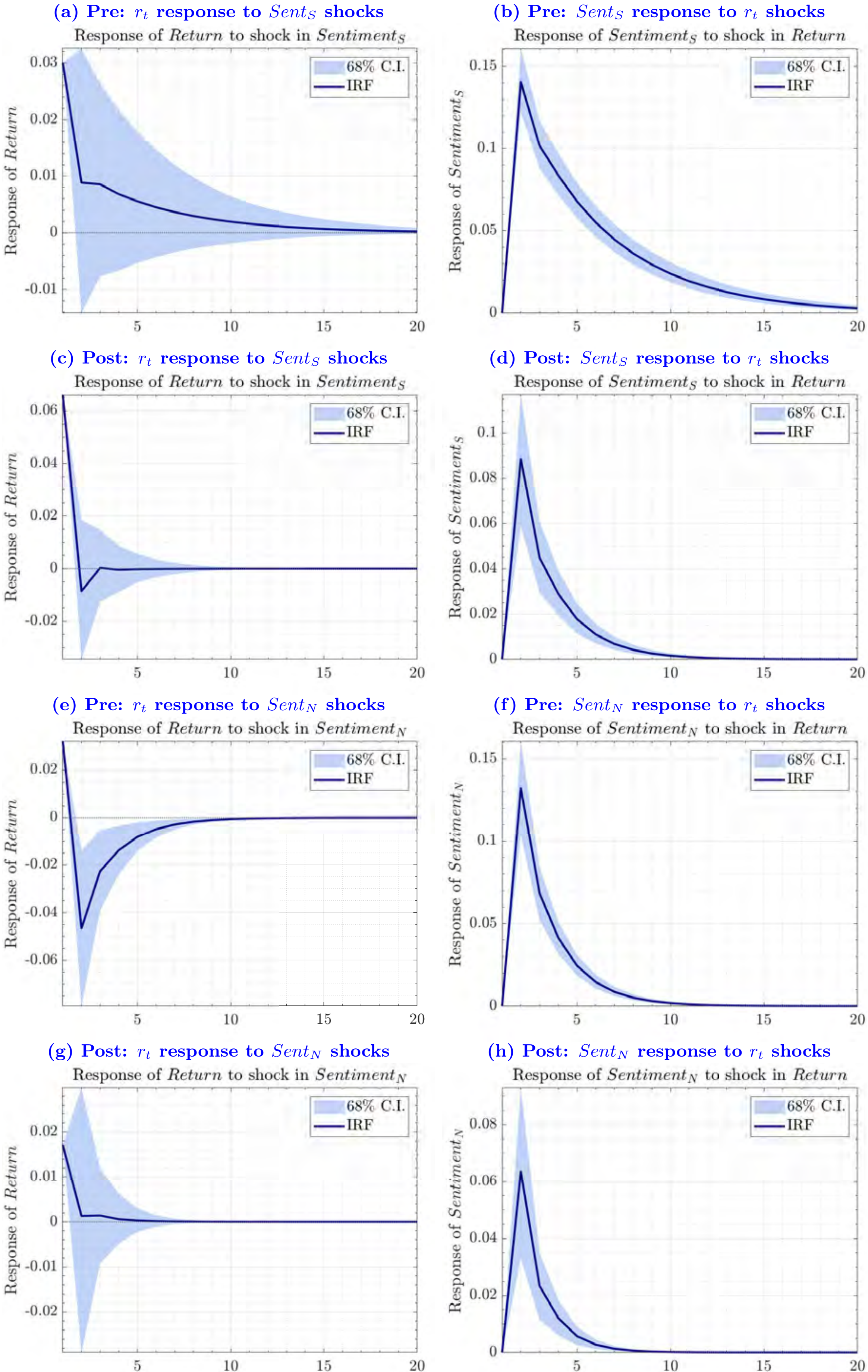
	Pre-transition Period					Post-transition Period			
ϕ_{11}	0.4469	0.0561	7.97***	0.00***	ϕ_{11}	0.4007	0.0572	7.00***	0.00***
ϕ_{12}	0.1060	0.0567	1.87*	0.06*	ϕ_{12}	0.2171	0.0949	2.29**	0.02**
ϕ_{21}	0.0849	0.0620	1.37	0.17	ϕ_{21}	0.0555	0.0374	1.48	0.14
ϕ_{22}	-0.0896	0.0626	-1.43	0.15	ϕ_{22}	-0.1257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

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

	Pre-transition Period					Post-transition Period			
	Value	SE	t-Stat	p-Value		Value	SE	t-Stat	p-Value
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			
	Value	SE	t-Stat	p-Value		Value	SE	t-Stat	p-Value
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***



(a) Pre: r_t response to $Sent_S$ shocks

(b) Pre: $Sent_S$ response to r_t shocks

(c) Post: r_t response to $Sent_S$ shocks

(d) Post: $Sent_S$ response to r_t shocks

(e) Pre: r_t response to $Sent_N$ shocks

(f) Pre: $Sent_N$ response to r_t shocks

(g) Post: r_t response to $Sent_N$ shocks

(h) Post: $Sent_N$ response to r_t shocks

Figure 4: *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).

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 panel (a) with panel (e) in Figure 4, we find that in the Pre-transition Period, return is 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, return over-corrects itself 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 panel (c) and panel (g) reveals that, at the Post-transition Period, return exhibits strikingly higher sensitivity to social media sentiment impact than to news sentiment, as manifest itself 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) than panel (g).

Panels of the IRFs on the right hand side of Figure 4 indicate the reverse causalities of each of its respective left hand side IRFs. All four panels (panel (b), (d), (f) and (h)) expose similar patterns: a unit of shocks from stock return causes positive and significant increases in both social media based and news based sentiment on the next day (spikes on 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 also has 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 costs 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 (panel (b), (d), (f) and (h) in Figure 4), we observe that both social media and news sentiment become less sensitive to return 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 panel (b) with (d), and comparing panel (f) with (h)), but in fact it is not. One possible explanation to this phenomenon could be resort to the scarcity of investor attention nowadays. The abundance in communication platforms and information channels facilitates information exchange 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 between return and sentiment in this subsection validate and extend the media induced structural transition patterns 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 innovations from the counterpart have both accelerated in the post-transition period compared with the pre-transition period, irrespective of the types of media that sentiment measure is based on. 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 in the post-transition session. Feedback effects from return to social media sentiment and to news-based sentiment, however, have both depreciated.

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-transition and post-transition periods. We estimate the following system equations, by representing $k = 2$, $x = (Sent^2_S, VIX)'$ and $x = (Sent^2_N, VIX)$ respectively into the General Setup.

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

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

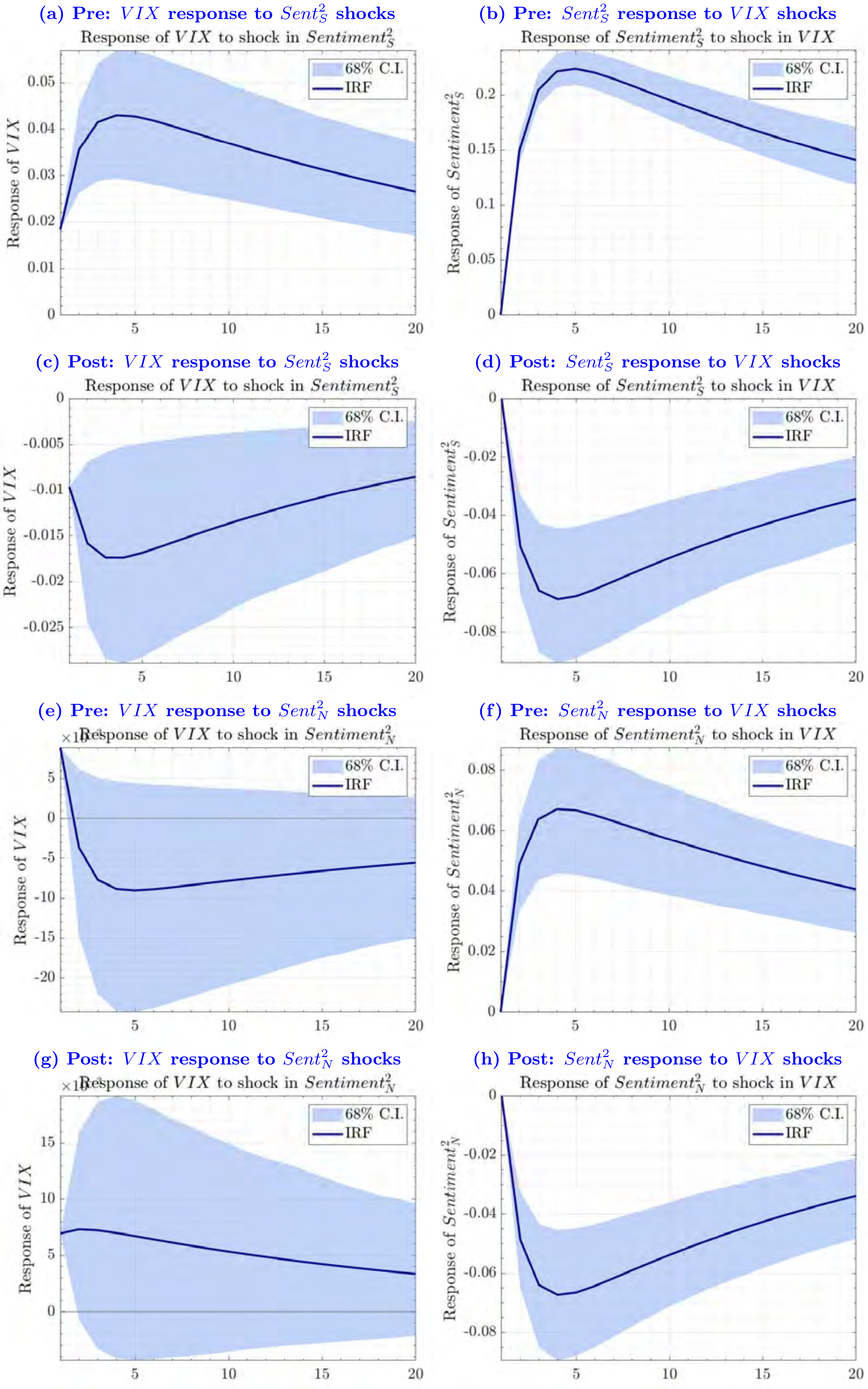
We choose VIX (V_t) as a measure of volatility in the above two systems because investor sentiment affects asset prices by shaping investors' beliefs about the future. In contrast, traditional realized volatility measures (RV), such as standard deviation or squared term of prior returns, are backward-looking. 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 ($Sent^2_S$ and $Sent^2_N$) as a measure of the high sentiment period with strong extreme emotions. The benefit of using squared term of sentiment lies in its incorporation of the disagreement of opinions expressed in social and news media. Since our sentiment scores are volume-weighted[23] 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 dominated by a similar kind of emotion, for example, most investors are extremely optimistic, or strongly angry, when observing $Sent^2$ close to 1. 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 Sen^2 might result from: i) weak emotions ex-

[23] 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.*

pressed in media; and ii) strong positive and negative emotions expressed but these parsed texts' scores cancelling with each other when forming the net sentiment value. 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 analysis performed 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.

Panel C and D in Table 6 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 autocorrelatioin effect is more salient than the cross-impacts between sentiment and volatility. However, these values are the initial responses only, which do not help us to 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 5 depicts the Impulse Response Functions (IRFs) of VIX responses to shocks from social media sentiment or news-based sentiment (left-hand side panels: panel (a), (c), (e) and (g)) in both the pre-transition 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 (panel (b), (d), (f), and (h)). The top two panels in both sides (panel (a), (b), (c), and (d)) are IRFs of the VIX and **social** sentiment VAR system, while the bottom two panels in both sides (panel (e), (f), (g) and (h)) are IRFs of the VIX and **news** media VAR system. In both panel (a) and (c), we find that VIX reaches its peak after 4 to 5 working days (about a week) following one unit of unexpected rises in social media sentiment (both positive and negative), and this process gradually corrects itself to 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 responces (left side IRFs in Figure 4), which all revert back to zero within our IRFs observation window, the reaction of volatility (left side IRFs in Figure 5) 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 volatility is negatively associated with the rising social media sentiment in the post-transition period (panel (c)). In contrast, VIX responses to news sentiment shocks exhibit totally different situations from social media. Comparing panel (e) with (g), we recognise similar levels (about 0.006 to 0.007) of initial VIX responses to news sentiment shocks in both the pre-transition and post-transition periods. The estimated IRFs coefficients do not fully revert back to zero after about a month - the same as the social media effects shown in panels (a) and (c). However, the broader error bands crossing zero in lagged one to two days after the shock indicate that volatility is less sensitive to news sentiment shocks than to social media sentiment shocks.



(a) Pre: VIX response to $Sent^2_S$ shocks

(b) Pre: $Sent^2_S$ response to VIX shocks

(c) Post: VIX response to $Sent^2_S$ shocks

(d) Post: $Sent^2_S$ response to VIX shocks

(e) Pre: VIX response to $Sent^2_N$ shocks

(f) Pre: $Sent^2_N$ response to VIX shocks

(g) Post: VIX response to $Sent^2_N$ shocks

(h) Post: $Sent^2_N$ response to VIX shocks

Figure 5: ***Sentiment*^{2} vs *VIX* Sub-sample Comparison.** Panel (a) to (d) are IRFs of $x_t = (Sent^2_S, V_t)'$; panel (e) to (h) are IRFs of $x_t = (Sent^2_N, V_t)'$. “Pre” denotes Pre-transition Period: 2011/01/01-2013/12/31; “Post” dentes Post-transtion Period: 2016/01/01-2017/11/30. Horizontal axis represent 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).

A comparison between the magnitudes of all four left hand side panels with the right hand panels in Figure 5 reflects the fact that the 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 (more significant in term of statistics) than their left-hand side counterparts. In the pre-transition period, the positive IRFs in panel (b) and (f) show that both social media and news-based sentiment spike higher following shocks from VIX, meaning less disagreement is found on media after VIX surges higher. In the post-transition period, however, the upward concaved IRFs in panel (d) and (h) illustrate that sentiment in social and news media, regardless of the media type, becomes more neutral and contains less disagreement of opinion: both IFRs plots touch the troughs (approximately -0.07) after approximately four working days.[24] In contrast to the fully correction situation in right side IRFs of Figure 4, none of the four right-hand side figures in Figure 5 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.

In sum, results from this subsection find consistent evidence to previous studies such as Araújo et al. (2018) that reverse causal effects from VIX to social/news media sentiment are stronger than the causal effects from media sentiment to volatility. We also find that VIX is more sensitive to social media sentiment than to news-based sentiment at both the pre-transition and post-transition sub-samples. Before 2014, the increased volatility is linked to a higher level of $Sent^2$ for both social media and news-based sentiment, indicating high volatility and strong emotions mutually cause each other in the pre-transition period. After 2016, the heightened volatility is associated with stronger extent in disagreement of opinion. A comparison between the analysis performed for return and sentiment systems in prior subsection (Figure 4) with analysis conducted in this subsection (Figure 5) 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. We find that before 2014, both the activeness (measured by *buzz*) and the emotions (measured by *sentiment*) expressed in news media significantly dominate those in social media. After 2016, however, both quantities and sentiment that appear in the social media play the leading role relative to news media. After identifying a period of structural transition in financial media landscape, we explore the dynamic lead-lag relationships between media sentiment and stock market return and volatility.

Our results suggest that media sentiment and stock variations mutually influence each other. Before 2014, return is more sensitive to news-based sentiment, while after 2016, return is more

[24] As stated in the previous paragraphs, We interpret the decrease in $Sent^2$ to a negative value (IRFs in panel (d) and (h)) as containing less disagreement in opinions rather than interpreting it as expressing weaker sentiment, because the emotional measures from TRMI is buzz-weighted, or have already controlled for the posting/coverage volumes.

responsive to social media induced sentiment. We find that the feedback effects from return to media sentiment expose a more salient pattern than the causal effect from sentiment to market. Most strikingly, we find that the speed of correction of market variations caused by sentiment has expedited after 2016, compared with the period before 2014.

Using squared term of media sentiment as a proxy for extreme emotions, we investigate the associations between volatility (VIX) and high level of sentiment intensity with respect to different types of media. Similar to the analysis between return and sentiment, We find that the impacts from VIX to media sentiment are more prominent than the effects from sentiment to VIX. VIX is more sensitive to social media sentiment than to news based sentiment at both pre-transition and post-transition sub-samples. The inter-relations between sentiment and VIX is more persistent than that of return.

Overall, this study offers three substantial contributions to the literature on investor sentiment and noise trader risk. First, our perspective of directly contrasting social media effects with news effects at different time periods echoes the importance of accounting for time-varying relationships between investor sentiment and stock market, as pointed out by Baker and Wurgler (2007). Second, modeling the dynamic influence between media sentiment and market variations, we help generate new insights to the field of textual analysis sentiment predictability. Last but not least, applying a novel type of dataset and performed detailed statistic analysis toward it, we contribute to the line of research that synthesize sentiment from multiple media sources. In general, our findings assist to shed light on how information is incorporated into stock prices and volatility with regard to the recent technological changes. We will delve into further analysis in this topic at individual firms' level and at a more granular frequency.

References

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.

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.

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.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements.

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.

Fehr, E. and Tyran, J.-R. (2005). Individual irrationality and aggregate outcomes. *Journal of Economic Perspectives*, 19(4):43–66.

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.

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.

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.

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

Jiao, P., Veiga, A., and Walther, A. (2018). Social media, news media and the stock market. *News Media and the Stock Market (September 25, 2018).*

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.

Le Bon, G. L. (1896). *The Crowd: A Study of the Popular Mind.* Classic Books Library.

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.

Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance*, 42(3):483–510.

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.

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

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.

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.

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.

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.

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

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

A Appendix

A.1 List of acronyms and notation

Table A.1: List of acronyms.

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
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&P 100	Standard & Poor's 100 Index
S&P 500	Standard & Poor's 500 Index
SEC	The US Securities and Exchange Commission
SIRCA	Securities Industry Research Centre of Asia-Pacific
SVAR	Structural 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
WSJ	The Wall Street Journal

A.2 Data sources and variable names

Table A.2: List of data sources and variable names.

Code/Symbol	Description
RIC	Reuters Identification Code
.SPY	RIC for SPDR S&P 500 ETF
CBOE	Chicago Board Options Exchange
Datastream	Thomson Reuters Datastream
MPTRXUS500	TRMI company group code approximating S&P 500 constituents
SIRCA	Securities Industry Research Centre of Asia-Pacific
WRDS	Wharton Research Data Services
$Buzz_N$	news media buzz (report volume in news media)
$Buzz_S$	social media buzz (posting volume in social media)
$Sent_N$	news media net sentiment (positive minus negative sentiment)
$Sent_S$	social media net sentiment (positive minus negative sentiment)
r_t	log return on day t
V_t	VIX (CBOE options volatility index) on day t

A.3 Optimal Lag Length Information Criterion

Table A.3: Optimal Lag Selection for VAR systems. Panel A tabulates AIC and BIC criteria from lag 1 to lag 12 for the VAR systems between social media *Buzz* and news *Buzz*. Panel B lists AIC and BIC till lag 12 for the VAR system between social media *Sentiment* and news *Sentiment*. Optimal lag is denoted with * and boldface. We believe BIC is more crucial to help make decision because it takes on Bayesian approach. AIC is also listed to facilitate judgment and for completeness. BIC of Panel A and B suggests 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

A.4 Why VAR(1) is Parsimonious Form VAR (7)

Table A.4: 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 are VAR model coefficients estimated as shown in the General Setup when $\mathbf{x} = (Buzz_S, Buzz_N)'$ and $p = 7$; right panel are coefficients estimated for the General Setup when $\mathbf{x} = (Sent_S, Sent_N)'$ and $p = 7$. p-Values illustrates that the inter-mediate lags' (lag 2 to lag 6's) coefficients are insignificant for both models, and most of the significant coefficients are concentrated on lag 1 and lag 7. Therefore, it indicates that VAR(1) might be a parsimonious form representation of VAR(7).

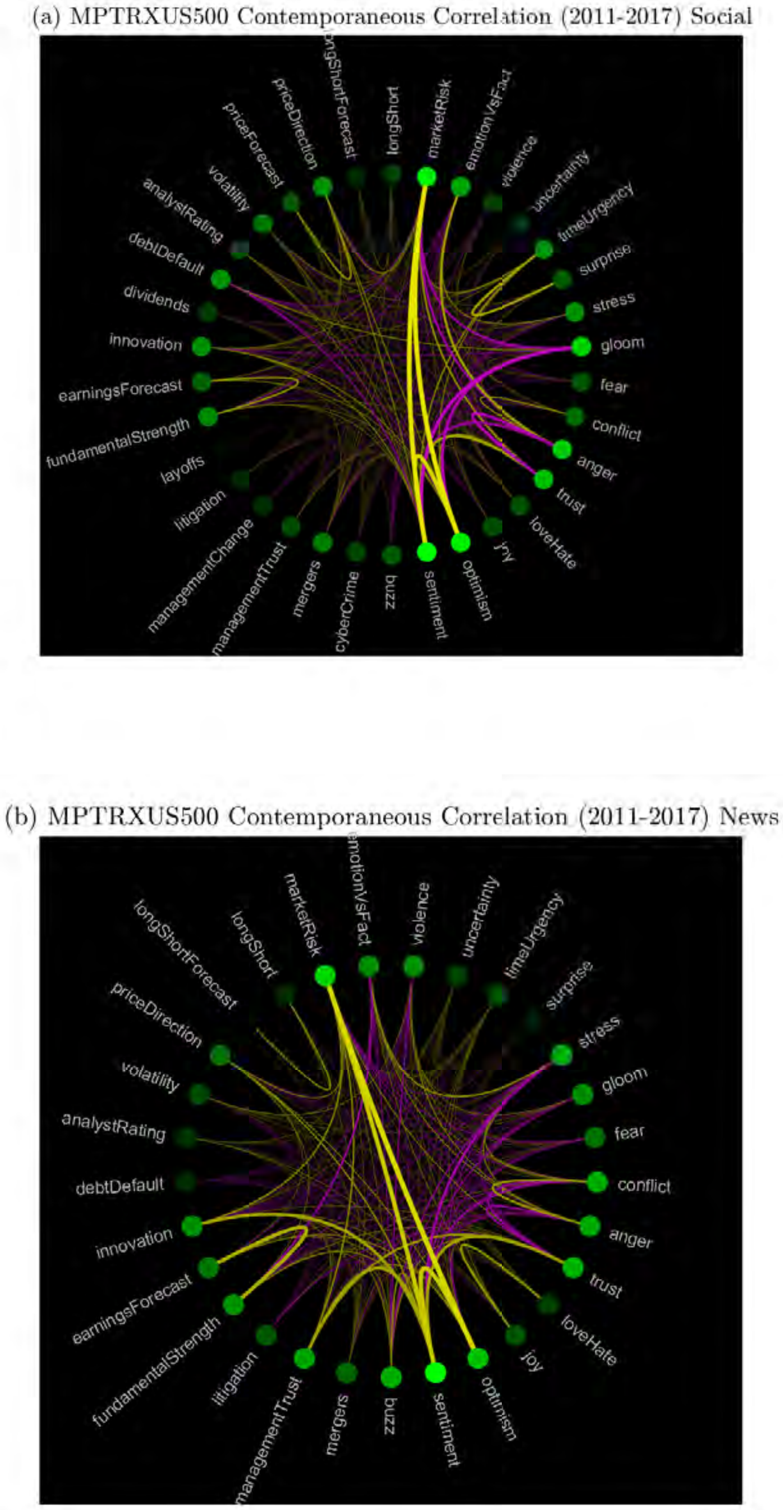
Sample A: First 365 days							
VAR(7): $Buzz_S$ **vs** $Buzz_N$				**VAR(7):** $Sent_S$ **vs** $Sent_N$			
	Value	SE	p-Value		Value	SE	p-Value
Constant1	-0.0742	0.0460	0.11	Constant1	-0.2109	0.0724	0.00***
Constant2	-0.2417	0.0694	0.00***	Constant2	0.0884	0.1284	0.49
AR1(1,1)	0.6312	0.0683	0.00***	AR1(1,1)	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 A.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 are VAR model coefficients estimated as shown in the General Setup when $\mathbf{x} = (Buzz_S, Buzz_N)'$ and $p = 7$; right panel are coefficients estimated for the General Setup when $\mathbf{x} = (Sent_S, Sent_N)'$ and $p = 7$. The results indicate 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 B: Last 365 days							
VAR(7): $Buzz_S$ **vs** $Buzz_N$				**VAR(7):** $Sent_S$ **vs** $Sent_N$			
	Value	SE	p-Value		Value	SE	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**

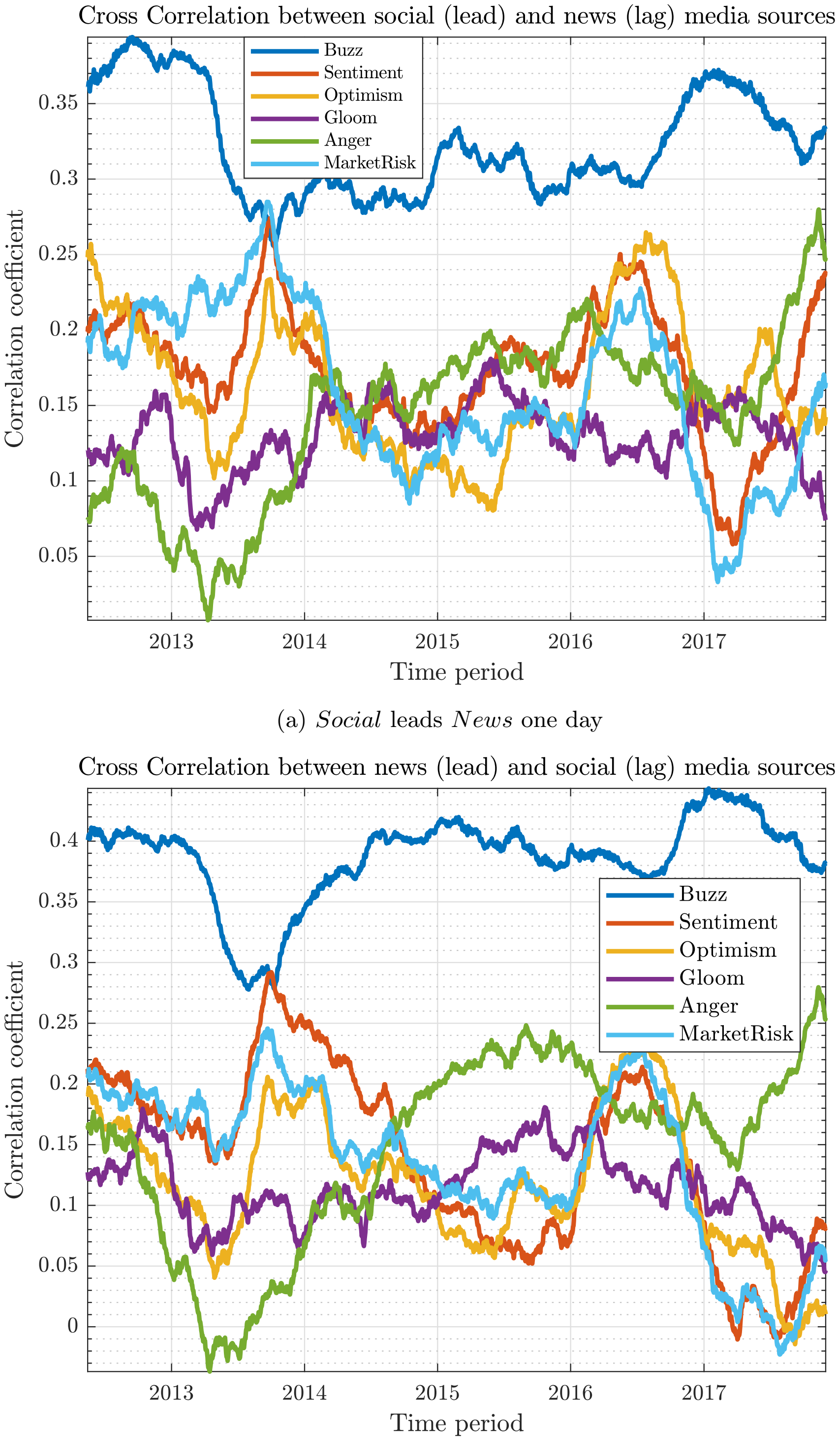
A.5 Correlation Schema-Balls


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

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

Figure A.1: Correlation coefficients between various emotional scores for the S&P 500 Company Group. The two panels are a visual representation of the pairwise contemporaneous correlations between all 35 scores for the S&P 500 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 2 and 3, 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.

A.6 One day lag cross correlations between social and news.

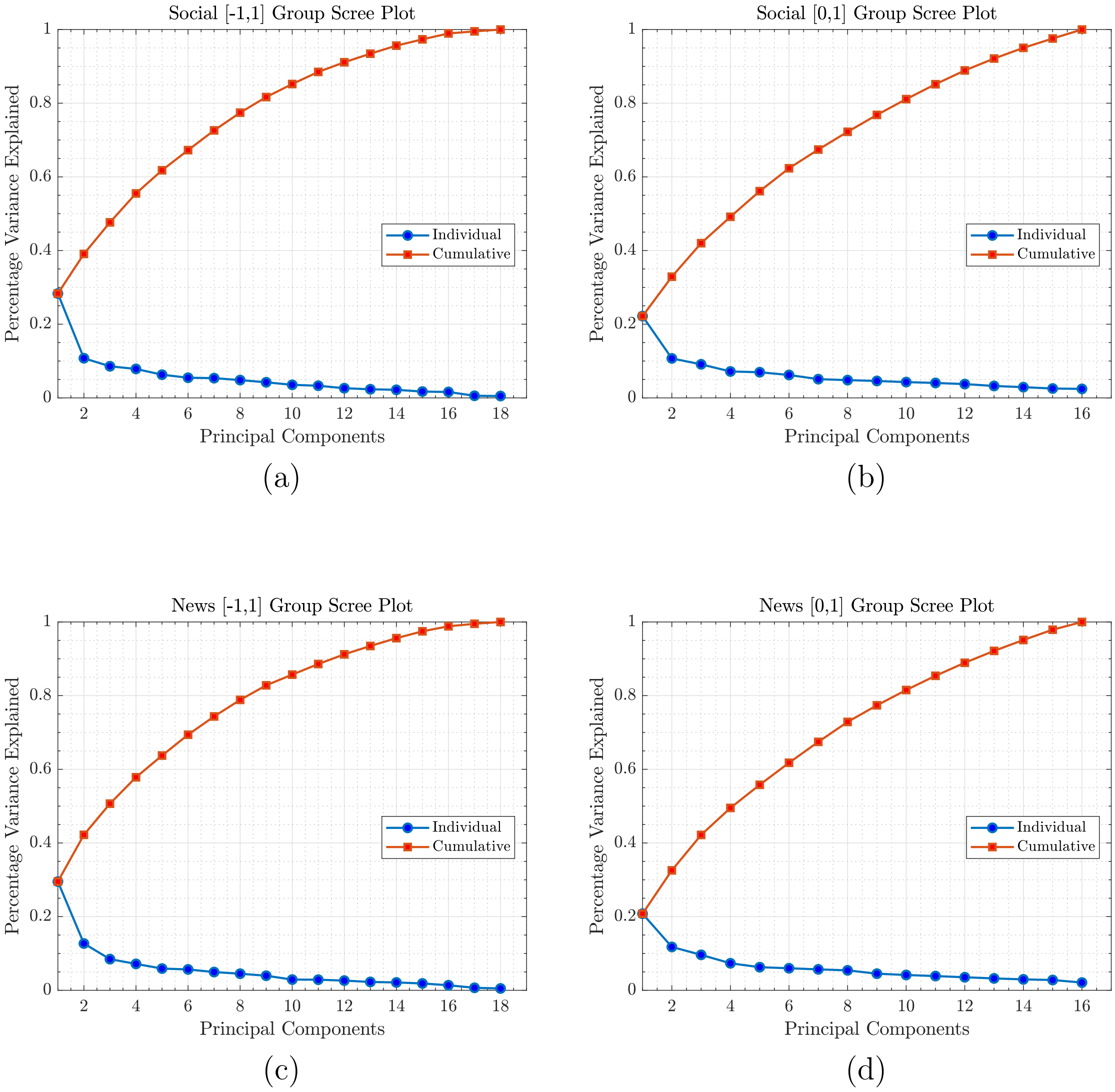


(a) *Social* leads *News* one day

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

Figure A.2: One Day Lag Cross-Correlation between S&P 500 Key Social and News scores. Panel (a) shows Kendal correlation between key social and news scores for for the S&P 500 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 1 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.

A.7 Scree Plots for Social and News Series



(a) (b) (c) (d)

Figure A.3: Scree Plots from Principal Component Analysis of emotional scores for the S&P 500 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. Panel (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.

A.8 Principal Component Analysis

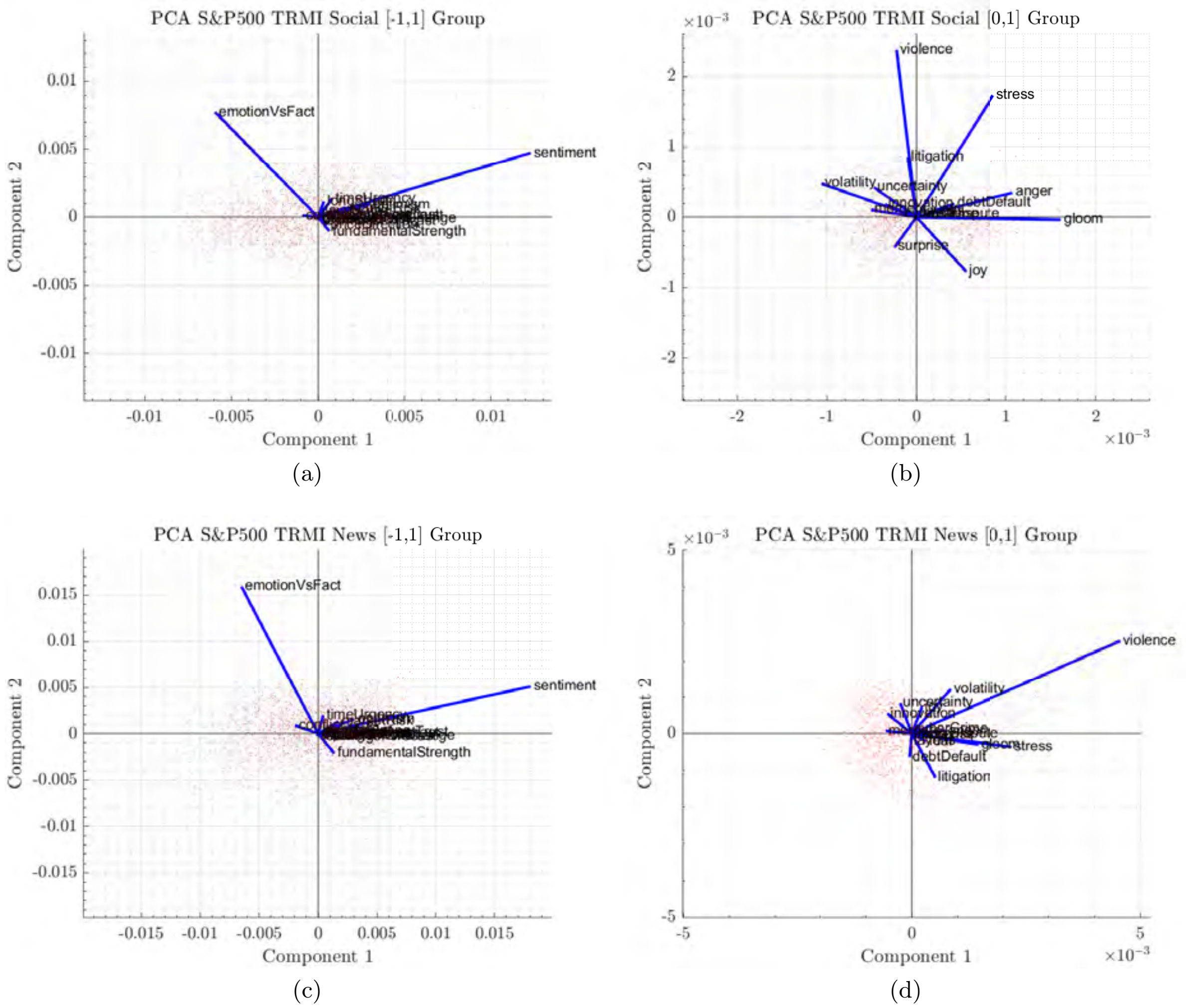


Figure A.4: Principal Component Analysis of the S&P 500 Sentiment Indices. Panel (a) is a biplot of the first two principal components for the [-1,1] sentiment score group in social sentiment indices; Panel (b) is a biplot of the first two principal components for the [0,1] sentiment score group in the social sentiment indices. Panel (c) and (d) are biplots constructed in a similar manner but using news sentiment data instead of social media. Panel (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 researches that focus 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.

A.9 ACF and PACF for main TRMI series

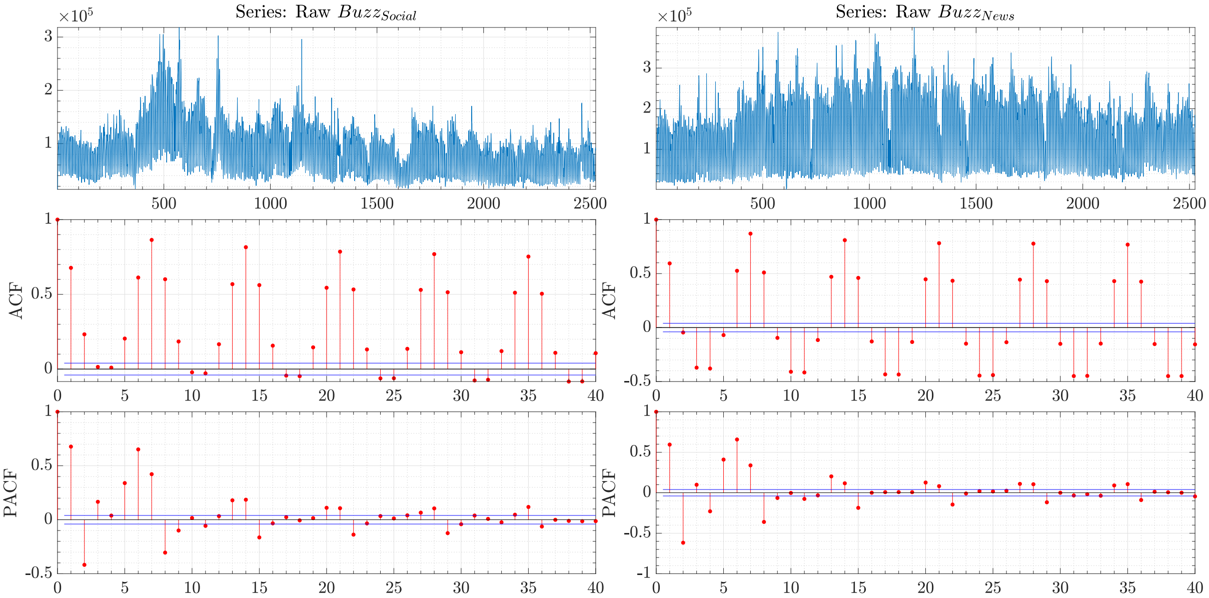


Figure A.5: Time-series Analysis of raw *Buzz* data. The left three panels show the sample distribution of the original social media posts volume measure: *Buzz*, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF) up to 40 days. The three panels on the right represent 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 2 and 3, 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 $Buzz_N$

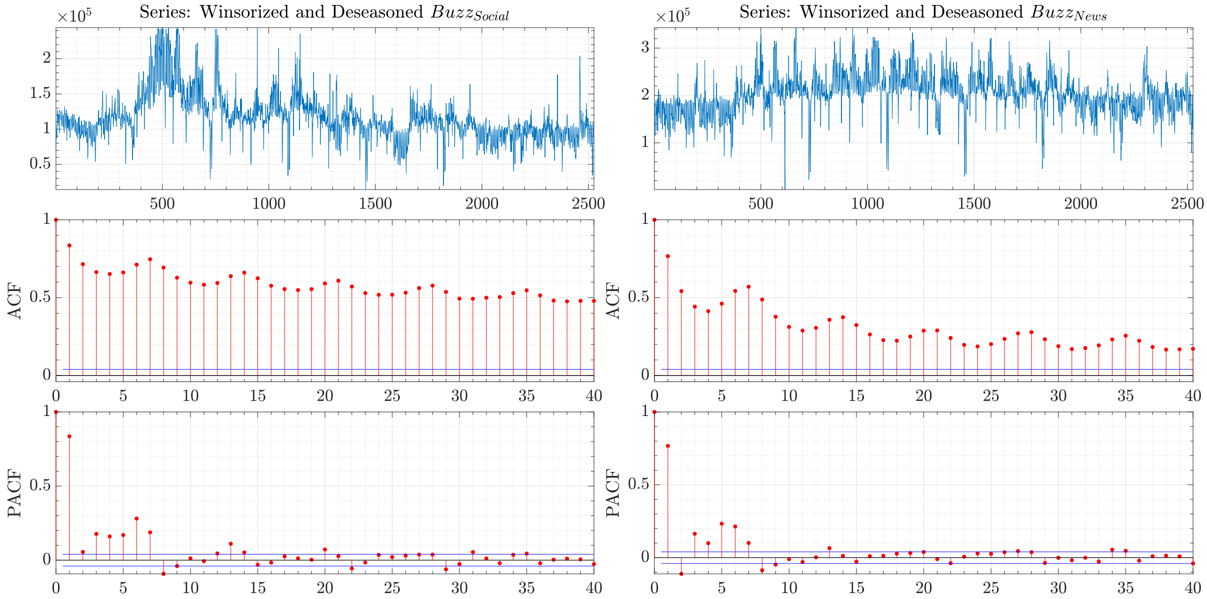


Figure A.6: Winsorized and De-Seasoned *Buzz* Time Series Check. The left three panels show the sample distribution of $Buzz_S$ after truncating the large value observations (asymmetric winsorizing the right tail outliers), Its autocorrelation function (ACF) and partial autocorrelation function (PACF) 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.5, 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.

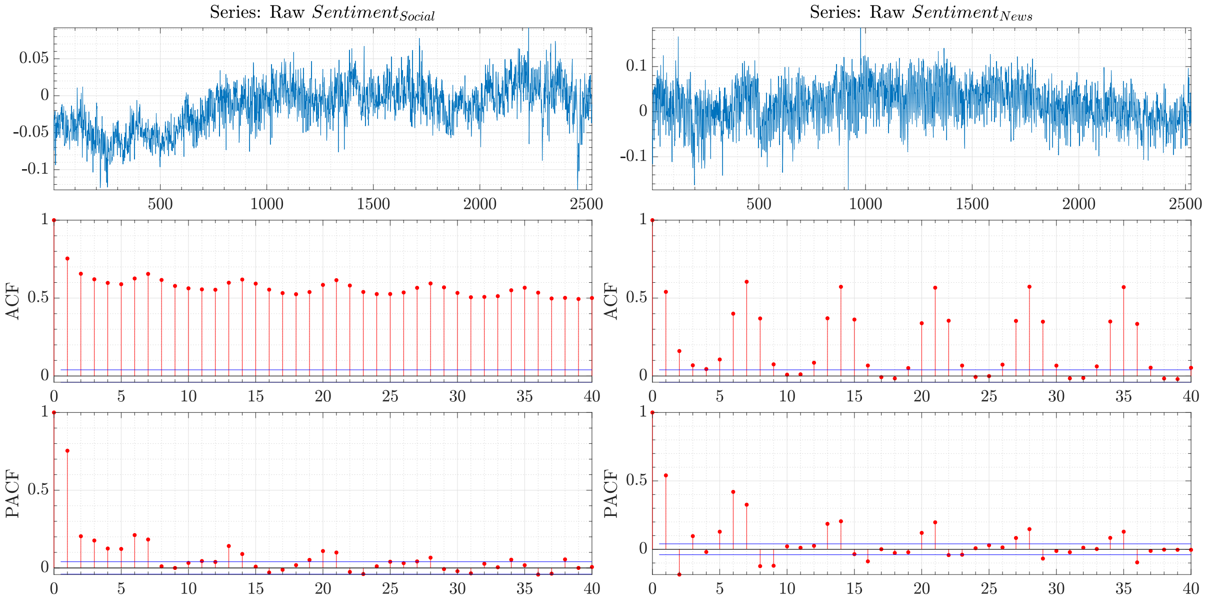


Figure A.7: Raw *Sentiment* Time Series Check. The left three panels show the sample distribution of the net positive and negative emotion scores from social media: $Sent_S$, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF) up to 40 days. The right side three 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 2 and 3. 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.

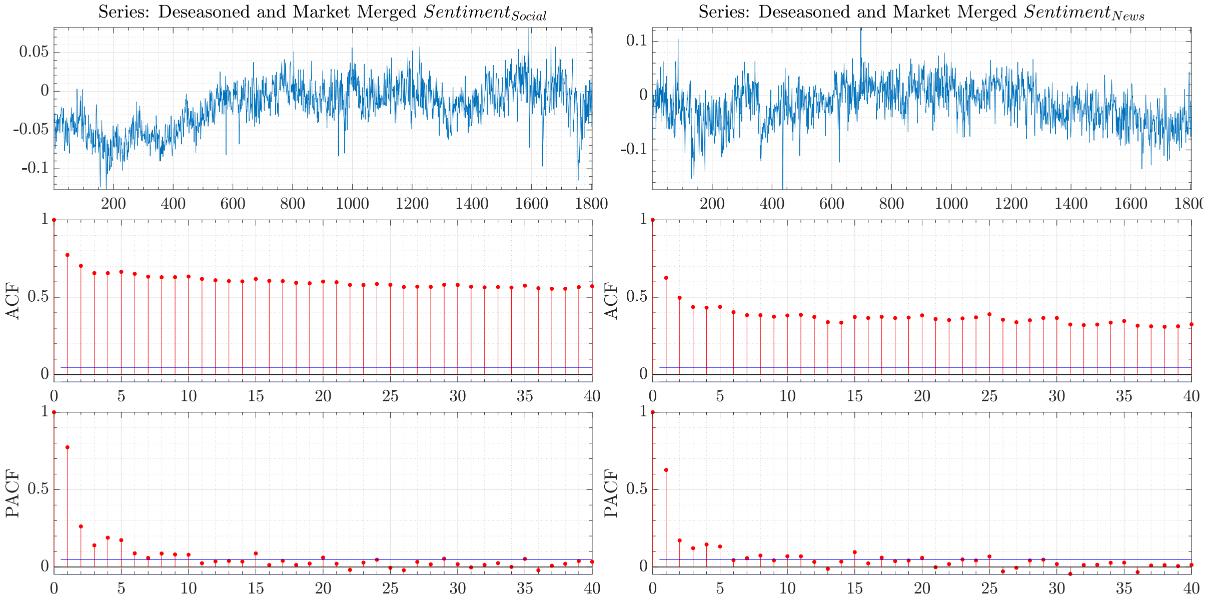


Figure A.8: De-Seasoned and Market Merged *Sentiment* Time Series Check. The left three panels show the sample distribution of the seasonality adjusted and non-trading day averaged value of $Sent_S$, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF) up to 40 days. 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.