

https://dspace.org/

Can a 280-character message explain stock returns?: evidence from the JSE

Nyakurukwa, Kingstone

2021
https://hdl.handle.net/10539/32586

Downloaded from a DSpace repository. https://dspace.org/



Can a 280-character message explain stock returns? Evidence from the JSE

Kingstone Nyakurukwa

School of Economics and Finance

University of the Witwatersrand

Johannesburg

South Africa



Can a 280-character message explain stock returns?
Evidence from the JSE

Kingstone Nyakurukwa
1483733
Supervisor: Dr.Y. Seetharam
School of Economics and Finance

A research report submitted to the School of Economics and Finance, Faculty of Commerce, Law and Management, The University of the Witwatersrand in fulfilment of the requirements for the degree of Master of Commerce (M.Com 50% research) in Finance.

Johannesburg, South Africa
February 2021

DECLARATION

I, Kingstone Nyakurukwa, declare that this research report is my own unaided work. It is submitted in fulfilment of the requirements for the degree of Master of Commerce (M.Com) in Finance at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university.

_____

Kingstone Nyakurukwa

February 2021

DEDICATION

To my late parents Kenyard and Beauty Nyakurukwa and my late brother Canaan Nyakurukwa

ACKNOWLEDGEMENTS

To my supervisor Dr Yudhvir Seetharam. I appreciate the supervision, mentorship and tutelage you have given me for the duration of this research project. Your inquisitive questions led me to the discovery of important academic nuggets I would not have discovered under normal circumstances. Your wealth of experience and exposure in the broad area of behavioural finance and asset pricing made life a lot easier for me.

To my wife Marifa M. Nyakurukwa and my daughter Candice K. Nyakurukwa. Your emotional support and encouragement helped me to cope with the many challenges I confronted during the period I was doing this research.

To my late parents Kenyard and Beauty Nyakurukwa. I will always cherish the belief and trust you have always shown in me. To my mum and brother Canaan, it has been a few months since you passed on and I know this would have made you proud.

To my siblings Cadwell and Gladys Nyakurukwa. You always give me a reason to push harder because of the belief you have in me.

To God be all the Glory.

Can a 280-character message explain stock returns?
Evidence from the JSE

ABSTRACT

Online stock forums have emerged as an essential investing platform where multiple users can share their opinions about financial markets. This study examines the association between *tweet* features (bullishness, message volume and investor agreement) and market features (stock returns, trading volume and volatility) using 158 South African companies and a dataset of firm-level Twitter and StockTwits messages extracted from Bloomberg for the period 1 January 2015 to 31 December 2019. Firstly, the results from the study show that there is a general contemporaneous association between *tweet* features and stock market features. Secondly, no monotonic relationship is found between the magnitude of *tweet* features and the magnitude of market features. The coefficients of *tweet* features are found to be stronger in absolute terms at the extreme quantiles of returns. The BDS tests confirm the nonlinear relationship between stock returns and tweet features which implies that the relationship should be modelled using nonlinear specifications. Thirdly, the study finds no evidence that past values of *tweet* features can predict forthcoming stock returns using daily data. However, analysis using weekly and monthly data shows that past values of *tweet* features contain useful content that can predict the future values of stock returns. This implies that policymakers must implement appropriate regulations to deter the development of bubbles or crashes during the cycles of *"greed"* and *"fear"*. The lack of causal effects between *tweet* features and market features at the daily frequency implies that the JSE is dominated by institutional investors who are not driven by sentiment. The findings from the study corroborate the findings from previous studies which confirmed the dynamic nature of efficiency on the JSE.

Keywords: Textual sentiment, Quantile regression, Twitter Sentiment, JSE
JEL Classification: G12, G14, G41

DEFINITION OF TERMS AND ABBREVIATIONS

API: Application Programming Interface. It is a software intermediation platform that allows two applications to interact. In textual analysis, it is used to extract texts from a microblogging platform.

Behavioural Portfolio Theory: This is a theory that postulates that investors form a portfolio that meets a broad range of goals. The portfolio can be thought of as a pyramid, where each layer represents a different goal.

Bots: Computer applications that are deployed on the internet to simulate human behaviour

Bullishness: This is the extent to which noise traders are hopeful or doubtful concerning the future. It is utilised as a proxy for investor sentiment in research that uses textual analysis to extract sentiment from texts.

Computational linguistics: Is a field in the engineering sciences that explores how human language can be automatically processed and interpreted.

Dumb money investors: Refers to retail investors who are often inexperienced and trade on sentiment.

EMH: Efficient Market Hypothesis. Is a hypothesis that posits that asset prices reflect all available information and therefore that it is impractical to beat the market.

The Fourth Industrial Revolution: refers to how technologies like artificial intelligence, autonomous vehicles and the internet of things are merging with humans' physical lives.

High-Frequency Traders: Are traders who utilise sophisticated algorithms to place large amounts of orders in fractions of a second. They are characterised by high speeds, high turnover rates and high order-to-trade ratios.

Investor Agreement: The extent to which investor opinions are unanimous about developments in the financial market. Proxies that have been used in previous studies include analyst forecasts and various agreement scores extracted from text messages.

LM: Loughran-McDonald dictionary – It is a finance-specific dictionary that classifies finance-related words into various sentiment categories.

Machine learning: This is the process of using sophisticated algorithms to analyse data, acquire knowledge from it and then make forecasts based on the knowledge acquired.

MarketPsych Index: Sentiment index that is derived from a distillation of large volumes of news and social media data. The index is derived and managed by Thomas Reuters Inc.

Microblog: A social media platform that allows the exchange of short messages among account holders.

POMS: Profile of Mood States – This is a five-point scale used by psychologists to measure transient, distinct mood states of an individual.

Quantile regression: This is a type of regression that estimates the association between the independent variables and the response variable across the distribution of the response variable. It extends the Ordinary Least Squares regression which only estimates the conditional mean of the dependent variable across values of the predictor variables.

Raging Bull: This is an online platform that permits users to post and respond to news and other finance-related news as well as participate in discussion rooms.

Robinhood: Is an American company that provides a mobile app and online platform that offers retail investors an avenue to invest in listed shares, ETFs and other financial products through its website and mobile apps without online fees. The investors who invest using this platform are called Robinhood investors.

Self Organising Fuzzy Neural Network: This is a learning machine that finds the parameters of a fuzzy system by utilising estimation methods from the field of neural networks.

Social media farms: Involves a group of low-paid workers who are recruited to generate likes, followers and replies on social media platforms.

StockTwits: Is a microblogging platform that allows investors, traders and entrepreneurs to share information and ideas concerning developments in the stock market.

Textual analysis: Is a methodology that attempts to understand the language and symbols present in texts to gain information on how people make sense of events.

Twitter: A microblogging and social networking platform on which account holders can post, reply and forward messages (also called *tweets*).

XPOMS: Extended Profile of Mood States – An extension of the POMS which includes some Africaans words.

Yahoo!Finance: Is an online platform that provides financial news on stock market quotes, press releases and financial reports.

TABLE OF CONTENTS

DECLARATION ..... **ii**
ACKNOWLEDGEMENTS ..... **iv**
ABSTRACT ..... **v**
DEFINITION OF TERMS AND ABBREVIATIONS ..... **vi**
TABLE OF CONTENTS ..... **ix**
LIST OF FIGURES ..... **xi**
LIST OF TABLES ..... **xii**
1 INTRODUCTION **1**
1.1 Background and overview 1
1.2 Knowledge gap 4
1.3 Problem statement 5
1.4 Research objective 5
1.5 Hypotheses 6
1.5.1 Primary hypothesis 6
1.5.2 Secondary hypothesis 6
1.5.3 Tertiary hypotheses 7
1.6 Summary of findings 7
2 LITERATURE REVIEW **10**
2.1 The Efficient Market Hypothesis 10
2.2 Behavioural Finance 11
2.3 Reconciling the EMH and behavioural finance 15
2.4 Social media and textual sentiment 17
2.5 Social media and wisdom of the crowds 20
2.6 The association between textual sentiment and stock market features 21
2.7 The asymmetric effects of sentiment on returns 25
2.8 Methods and models used for textual sentiment examination in finance 29
2.8.1 Sources of information 29
2.8.2 Content analysis methods 31
2.8.3 Measuring textual sentiment 33
2.8.4 Econometric models 33
2.9 Summary 35
3 DATA AND RESEARCH METHODOLOGY **37**
3.1 Data 37
3.2 Variables 37
3.2.1 *Tweet* Features 37
3.2.2 Market Features 40
3.3 Econometric model estimation 42
3.3.1 Contemporaneous relationship between *tweet* features and stock market features 42
3.3.2 Pooled quantile regressions 42
3.3.3 Granger non-causality tests 44

3.4 Robustness checks 45
3.4.1 Random effects regression 46
3.4.2 Quantile regressions with fixed effects 46
3.4.3 Lag-lead relationships 47
3.5 Summary 48
4 RESULTS AND DISCUSSION 49
4.1 Descriptive statistics 49
4.2 Contemporaneous relationship between *tweet* features and market features 53
4.3 The relationship between the magnitude of stock returns and *tweet* features 56
4.3.1 Bullishness and stock returns 57
4.3.2 Investor agreement and stock returns 59
4.3.3 Message volume and stock returns 61
4.3.4 Equality of coefficients 63
4.4 The information content of past values of *tweet* features 65
4.5 Robustness Checks 68
4.5.1 Quantile regressions using alternative conditional quantiles 68
4.5.2 Fama and Macbeth style regressions 70
4.5.3 Granger non-causality test using weekly and monthly data 72
4.5.4 Brock-Dechert-Scheinkman (BDS) tests for non-linearity 76
5 CONCLUSION 79
REFERENCES 82
Appendix A: List of companies 88
Appendix B: Panel unit root tests 90
Appendix C: Robustness tests 93
Appendix D: R-code for Quantile regression 102

LIST OF FIGURES

Figure 1: Number of social media users in South Africa according to Statista 18
Figure 2: Distribution of *tweets* by day of the week 50
Figure 3: Daily distribution of the aggregate *tweet*s 51
Figure 4: Bullishness - QR and OLS estimates with the 95% confidence intervals 58
Figure 5: Agreement - QR and OLS estimates with the 95% confidence intervals 60
Figure 6: Message volume - QR and OLS estimates with the 95% confidence intervals 62
Figure 7: Visualising quantile estimators using $\tau = \epsilon[0.1: 0.9]$ 69
Figure C1: Visualisation of the quantile estimators using $\tau=\epsilon[0.05,0.2,0.4,0.6,0.8,0.95]$ 101

LIST OF TABLES

Table 1: Summary statistics of the variables 49
Table 2: Correlations among tweet features and market features 52
Table 3: Results from the Hausman test 53
Table 4: Contemporaneous OLS regressions with firm fixed effects 54
Table 5: Pooled quantile regression estimates 56
Table 6: Quantile slope equality test 64
Table 7: Summary of panel unit root tests 66
Table 8: Pairwise Dumitrescu Hurlin Panel Causality Tests 67
Table 9: Results from quantile regressions using $\tau = \epsilon[0.1: 0.9]$ 68
Table 10: Lead-lag Fama Macbeth style regressions of returns on tweet features 71
Table 11: Lead-lag Fama Macbeth style regressions of returns on tweet features 72
Table 12: Panel unit root tests using weekly and monthly data 74
Table 13: Pairwise Granger causality tests using weekly data 75
Table 14: Pairwise Granger causality tests using monthly data 76
Table 15: BDS statistics on residuals from regression of returns on tweet features 77
Table 16: BDS statistics on residuals recovered from GARCH (1,1) fitted on stock returns 78
Table A1: List of companies included in the study 88
Table B1: Eviews output from unit root tests using daily data 90
Table B2: Eviews output from unit root tests using weekly data 91
Table B3: Eviews output from unit root tests using monthly data 92
Table C1: Random effects estimation 93
Table C2: Quantile Regression with firm-fixed effects 93
Table C3: 1-day lag Granger non-causality tests 94
Table C4: 2-day lag Granger non-causality test 95
Table C5: 1-week lag Granger non-causality tests 96
Table C6: 2-week lag Granger non-causality tests 97
Table C7: 1-month lag Granger non-causality tests 98
Table C8: 2-month lag Granger non-causality tests 99
Table C9: Quantile estimators using $\boldsymbol{\tau} = \epsilon[0.05, 0.2, 0.4, 0.6, 0.8, 0.95]$ 100

1 INTRODUCTION

"These websites [Facebook and Twitter] can provide perhaps the most powerful mechanisms available to a private citizen to make his or her voice heard."

—U.S. Supreme Court Associate Justice Anthony M. Kennedy
Packingham v. North Carolina 582 US 8 (2017)

1.1 Background and overview

One of the herculean tasks that have preoccupied academics and investment practitioners for decades has been the need to comprehend the factors that are critical in pricing financial assets. Elementary research on the pricing of financial assets postulated that prices follow a stochastic process and this was the main thrust of the Efficient Market Hypothesis (EMH) (Fama, 1965). The EMH hypothesises that asset prices capture all available (and relevant) information and that prices react only when new information is filtered into the financial markets. Since the arrival of new information cannot be predicted, it therefore follows that asset prices are unpredictable (Fama, 1991).

According to classical finance, financial markets are informationally efficient[1] and therefore it is not feasible to earn above risk-adjusted returns. In their study, De Long, Shleifer, Summers and Waldmann (1990) show that rational arbitrageurs are always available to push asset prices back to their fundamental values by counteracting the influence of irrational investors. However, contrary to the EMH, the Limited Arbitrage Hypothesis argues that perfect arbitrage is not possible in the real world as arbitrageurs may find it difficult to drive prices to their fundamental values since all arbitrage activity is risky and requires capital to be effective (Shleifer & Vishny, 1997). The history of the stock market is inundated with some striking events which traditional asset pricing models have failed to explain. A more recent body of literature has ascribed some of the stock market anomalies observed over time to the noise created by trades which are driven by investor sentiment (Black, 1986; Baker & Wurgler, 2006). Kearney and Liu (2014) categorise sentiment that can affect stock markets into two groups: investor sentiment – where opinions concerning future cashflows and accompanying risks are not commensurate with the available facts; and textual

[1] An **informationally efficient market** is one in which all information concerning a firm's stock has been captured into its current price.

sentiment – which demonstrates the extent of optimism and pessimism in texts. The growth of textual sentiment analysis has mainly been driven by the rise of stock microblogs[2].

The importance of microblogs in the capital markets can be understood from the events that took place on 23 April 2013 when hackers hacked into the Twitter account of the Associated Press. A false *tweet* to the effect that the President of the United States of America, Barack Obama, had been wounded from detonations at the White House wiped out $136 billion of the market capitalisation of the S&P 500 Index in seconds (Matthews, 2013). Another example is the plunge of the Tesla Inc. stock by as much as 13%, wiping out US$13 billion of market value in a few minutes after the Chief Executive Officer of the company, Elon Musk, tweeted to his more than 33 million followers that the Tesla stock was too high (Higgins, 2020). More recently, the role of social media in the financial markets has been brought to the fore by the events surrounding Gamestop Inc., a struggling, Texas-based video game retailer. Wall Street hedge funds placed major bets on the demise of the Gamestop stock because of the effects of the COVID-19 pandemic on the brick-and-mortar company. Some retail investors who share views on Reddit, an online discussion forum, rallied together to buy the Gamestop stock. These retail investors were not buying the stock because of its fundamentals but because of the sentiment attached to the stock since most of them grew up using the company's video stores. The Gamestop stock soared by more than 1700% as a result, triggering a short squeze on several hedge funds who had taken short positions in the stock.

The aforementioned examples offer a live demonstration of how financial markets can respond to information that comes from microblogs. Also, the advent of high-frequency traders has revolutionised how trades are executed. High-frequency traders apply algorithms to scrutinise the activity on the internet in real-time by *'reading'* news and microblog messages and transacting on said information. It would be of interest to see the impact of this reaction on emerging market stocks. This study, therefore, uses textual sentiment contained in the 280 characters of Twitter and StockTwits individual messages to understand how they relate to stock market features.

[2] The terms "Microblogs", "Twitter" and "StockTwits" will be used interchangeably for the remainder of the study.

Although Shleifer and Vishny (1997) demonstrate that investor sentiment has a symmetric impact on asset returns, Prospect Theory shows that a *"dumb money"* investor is more likely to feel more discomfort from a loss than pleasure from a gain of comparable magnitude (Kahneman & Tversky, 1979). This loss aversion trait leads to investors having substantially different behaviours during poor market condition periods compared to favourable market condition periods. Thus, Prospect Theory predicts that the predictability of the stock market is likely to be pronounced and significant during periods with poor market conditions. On the other hand, Lemmon and Portniaguina (2006) postulate that naive investors are more likely to enter the stock market following bullish periods than bearish times, leading to more pronounced predictability of returns during good market conditions. This study investigates whether *tweet* features[3] have a symmetric effect on market features by establishing if there is a monotonic link between the magnitudes of the *tweet* features and market features.

Twitter and StockTwits platforms have been chosen for this study ahead of other microblogs because they are some of the most used by the investing community. These platforms allow participants to tap into the *"wisdom of crowds"*, where the sum of information emanating from multiple novice investors is presumed to predict outcomes more accurately than experts (Bartov, Faurel & Mohanram, 2018). Also, the short format of the platforms (up to 280 characters) and ease of information search (for example the use of *cashtags*, which are stock ticker symbols prefixed with a dollar sign), make them the ideal media to share information promptly and therefore relevant for a study of this nature. One of the pioneering studies that examined the role of sentiment extracted from *tweets* in the financial markets (Bollen, Mao & Zeng, 2011) led to the formation of the world's first Twitter-based hedge fund (Thompson, 2011). A study of this nature, therefore, helps to clarify if investment strategies based on social media are feasible in an emerging economy like South Africa. The choice of the JSE is premised on its globally recognised standard of regulation as well as its often dual classification as both emerging and developing (Seetharam, 2021).

[3] *Tweet* features refer to bullishness, message volume and investor agreement. These are explained in detail in Section 3.2.1.

Studies done to establish the relationship between textual sentiment extracted from internet-based texts as well as media-expressed texts and stock returns have mostly confirmed the strong predictive power of sentiment on share returns. Findings from these studies show that an increase in negative sentiment results in a fall in share prices (Tetlock, 2007); negative sentiment is negatively associated with abnormal returns (Antweiler & Frank, 2004) and that there is an increase in abnormal returns caused by an increase in the positivity of textual sentiment contained in press releases (Hillert, Jacobs & Müller, 2018). On the relationship between textual sentiment and volatility, findings from studies show that textual sentiment helps to predict volatility (Antweiler & Frank, 2004); there is an inverse association between a pessimism factor and the conditional volatility of the Dow Jones Industrial Index (Tetlock, 2007) and that a large amount of uncertain words leads to higher subsequent return volatility (Loughran & Mcdonald, 2011).

Prior studies have also examined the role of investor disagreement on stock return and trading volume. Classical finance theory hypothesises that increased disagreement leads to increased trading volume emanating from two participants giving divergent values to a financial asset (Karpoff, 1986; Kim & Verrecchia, 1991; Harris & Raviv, 1993). This theory has been empirically reinforced by the findings of Antweiler and Frank (2004), in which disagreement among microbloggers is linked with increased trading volume and stock returns. Motivated by the evidence from the aforementioned existing studies, the present study examines the association between *tweet* features and stock market features. The study uses companies constituting the FTSE/JSE All Share Index (JALSH) between 1 January 2015 and 31 December 2019.

1.2 Knowledge gap

Textual analysis is an emerging method of extracting investor sentiment which has not been extensively researched in emerging and frontier markets. Several studies have been done in South Africa that sought to establish the nexus between investor sentiment and a host of dependent variables like stock market returns (Dalika & Seetharam, 2015); return volatility (Rupande, Muguto & Muzindutsi, 2019) and foreign financial flows (Muguto, Rupande & Muzindutsi, 2019) among others. However, these studies have mainly used survey-based investor sentiment which is usually low frequency and does not capture investor sentiment in real-time. This study seeks to augment this area by using investor sentiment extracted from high-frequency data from Twitter

and StockTwits. Johnston and Maree (2015) extracted sentiment from high-frequency data from Twitter but they however only used a sample period spanning 55 days. This study complements the body of knowledge since it uses a longer sample period (1247 trading days) and uses firm-level *tweet* features compared to the market-level *tweet* features used by Johnston and Maree (2015).

1.3 Problem statement

Proponents of the EMH agree that financial asset prices follow a random walk and cannot, therefore, be predicted because of high levels of noise (Malkiel, 2003). An emerging crop of behavioural finance researchers starting with Shiller (1981), however, argue that financial markets are predictable since people often make irrational decisions, contrary to assumptions of classical finance theory. Though studies have been done to validate the use of technical, fundamental and textual analysis to predict asset prices, empirical results from these studies have been mixed, which makes the validity of these techniques largely disputed. The present study is more comprehensive and robust as it examines the relationship between sentiment and market features at all quantiles of the distribution of the market features. Additionally, the increased utilisation of social media by companies and the investing community has increased calls for studies that investigate the predictive power of textual sentiment. This study attempts to fill these gaps by examining if textual sentiment from social media can explain stock market features in a South African context.

1.4 Research objective

The main objective of the study is to establish if *tweet features* (bullishness, message volume and investor agreement) can explain a cross-section of market features (stock return, trading volume and volatility). The study aims to specifically:

- Determine if firm-level *tweet* features are contemporaneously associated with stock market features;
- Determine if the magnitude of *tweet* features is monotonically related to the magnitude of stock market features;
- Determine if past values of *tweet* features contain useful information that could be used to predict future values of stock returns.

1.5 Hypotheses

1.5.1 Primary hypothesis

Determine if firm-level *tweet* features are contemporaneously associated with market features;

$H_{0,A}$: *Firm-level tweet features are not contemporaneously associated with stock returns.*
$H_{1,A}$: *Firm-level tweet features are contemporaneously associated with stock returns.*

$H_{0,B}$: *Firm-level tweet features are not contemporaneously associated with trading volume.*
$H_{1,B}$: *Firm-level tweet features are contemporaneously associated with trading volume.*

$H_{0,C}$: *Firm-level tweet features are not contemporaneously associated with volatility.*
$H_{1,C}$: *Firm-level tweet features are contemporaneously associated with volatility.*

1.5.2 Secondary hypothesis

Determine if the magnitude of *tweet* features is monotonically related to the magnitude of market features:

$H_{0,D}$: *The magnitude of firm-level tweet features is not monotonically related to the magnitude of stock returns.*
$H_{1,D}$: *The magnitude of firm-level tweet features is monotonically related to the magnitude of stock returns.*

$H_{0,E}$: *The magnitude of firm-level tweet features is not monotonically related to the magnitude of trading volume.*
$H_{1,E}$: *The magnitude of firm-level tweet features is monotonically related to the magnitude of trading volume.*

$H_{0,F}$: *The magnitude of firm-level tweet features is not monotonically related to the magnitude of volatility.*

$H_{1,F}$: *The magnitude of firm-level tweet features is monotonically related to the magnitude of volatility.*

1.5.3 Tertiary hypotheses

Determine if past values of *tweet* features contain useful information that could be used to predict future values of returns.

$H_{0,G}$: *Past values of tweet features do not contain useful information that could be used to predict future returns.*

$H_{1,G}$: *Past values of tweet features contain useful information that could be used to predict future returns.*

1.6 Summary of findings

The study examines whether information from stock microblogs (Twitter and StockTwits) can explain stock market features. To this end, three questions are asked to know the role that opinions from the Twitter and StockTwits platforms play in determining what happens in the stock market: Is there a contemporaneous association between *tweet* features and market features? Is the magnitude of *tweet* features monotonically related to the magnitude of market features? Do the past values of *tweet* features contain important information that could be used to predict the market features. Concerning the first research question, the results largely show a statistically significant contemporaneous relationship between *tweet* features and stock market features. The only associations which are not statistically significant at the conventional levels of significance are the associations between message volume and returns as well as between bullishness and trading volume.

While most previous studies on investor sentiment have used mean-based estimators like the Ordinary Least Squares (OLS) specifications to examine the association between *tweet* features and market features, these methods have been recently criticised as they assume symmetric associations between the variables. This study examines the relationship between the magnitude

of *tweet* features and market features across the distribution of the latter rather than the mean only. This way, it is possible to establish whether a link between the variables is symmetric. Using a relatively new but extensively used econometric methodology in recent times, quantile regressions, the study finds an asymmetric and non-monotonic relationship between *tweet* features and market features. The absolute magnitudes of the relationship between *tweet* features and market features are found to be trivial at mid-quantiles of the market features while significant at the extreme tails of the distributions of the market features. This implies that policymakers have to implement appropriate policies to deter the formation of bubbles or crashes during the greed and fear cycles.

While the findings outlined above are important in trying to understand the role of microblogs in the financial markets, a typical investor would be interested in knowing whether past values of *tweet* features contain useful information that could be used to predict future values of stock returns. The results from this study, using panel Granger non-causality tests and lagged Fama and Macbeth-style regressions show a non-causal relationship from *tweet* features to stock market features using one-day and two-day lags. However, using lower-frequency (weekly and monthly) granularity analysis, the results show that past values of *tweet* features contain important information that could be used to predict future returns. The predictive power of *tweet* features on stock returns at daily intervals but not at weekly and monthly intervals shows the dynamic nature of market efficiency on the JSE. On the other hand, using daily, weekly and monthly granularity analysis, a causal relationship from stock returns to market features is established. This shows that irrespective of the frequency of the data, lagged values of stock returns contain useful information that can be used to forecast future values of *tweet* features.

The study proceeds as follows. Section 2 reviews the existing literature on the theoretical framework underpinning this study as well as empirical studies showing the role played by investor sentiment, particularly from online texts, on the stock market. The section also includes a survey of methodological issues surrounding textual analysis. This is followed by Section 3 which outlines the econometric models used to test the hypotheses developed for this study. Section 4 presents the empirical results from the econometric and statistical specifications as well as various robustness checks to establish if the results are sensitive to methodological considerations. The final section, Section 5, summarises the findings and presents the conclusions of the study. In this

section, recommendations about the direction future studies could take to broaden and deepen this study are also proposed.

2 LITERATURE REVIEW

"The efficient market hypothesis is the most remarkable error in the history of economic theory."
- Lawrence Summers, U.S. Secretary of Treasury, on the Black Monday crash of 1987

This segment provides a critical desktop review of the existing studies in the research area. The theoretical framework is outlined first and includes a discussion of the EMH and the emergence of the behavioural finance paradigm as an alternative to some of the market anomalies characterising the stock market. The section ends with a discussion of a survey of empirical literature that looks at the link between *tweet* features and market features as well as a survey of methodological considerations relating to the topic.

2.1 The Efficient Market Hypothesis

The idea that markets are efficient was first documented in the 19th Century as some of the earliest studies reported that the prices of government bonds followed a random walk (such as Bachelier, 1901). In 1933, Cowles corroborates this earlier research by arguing that it is impossible to predict asset prices accurately (Cowles, 1933). Kendall (1953) provocatively introduces his *"demon of chance"* metaphor where he argued that asset prices are stochastic and the next week's prices do not depend on what the prices would have been the previous week as investors only depend on a blind chance. The idea of efficient markets was first formalised by the independent work of Samuelson (1965) and Fama (1965, 1970).

Long considered the cornerstone of modern financial theory, the EMH proclaims that capital markets are informationally efficient as they instantaneously assimilate all available information (Fama, 1970). According to this hypothesis, the arrival of new information in capital markets leads to the prompt correction of the prices of stocks to their *'correct values'* (Malkiel, 2003). The EMH is closely related to the notion of a stochastic process which postulates that asset prices do not have a memory and that future price changes represent random deviations from prior prices (Malkiel, 2003). By definition, the coming of new information is presumed to be a chance event and as such, subsequent prices are also anticipated to be erratic and unpredictable. This characteristic of share

prices, in principle, means that it is not feasible to attain returns that exceed risk-adjusted market returns.

Fama (1970)'s definition of efficiency can be summarised by the three levels of efficiency which he formulated; *weak-form*, *semi-strong form* and *strong-form* efficiency. The weak-form version of efficiency posits that the current price of a stock reflects all the past information of the stock price and that utilising technical analysis to study the patterns of past stock prices to gain riskless gains is futile. The semi-strong version of market efficiency asserts that all publicly available information is already incorporated in security prices to an extent that it is impossible to earn above the risk-adjusted excess returns using technical analysis and/or fundamental analysis. Finally, the strong-form efficiency is the pinnacle of market efficiency where all information, public and private, is instantaneously incorporated in a stock price such that even insider trading cannot lead to returns above risk-adjusted excess returns.

Though the EMH has been widely accepted by the academic fraternity, it has failed to explain some stock market-related events that have taken place in the course of the history of capital markets. Some of these stock market-related events include the 1987 *"stock market crash"*, the *"internet bubble"* of 2000 and the *"sub-prime mortgage crisis"* of 2008. The failure by the EMH to explain some of the above events and other market anomalies like the excess volatility, the January effect and the day of the week anomaly contributed to the emergence of behavioural economics as a new paradigm that arose as an alternative explanation for these irregularities.

2.2 Behavioural Finance

In the 1980s, many economists started questioning the EMH by arguing that stock prices are somewhat predictable. These researchers emphasised the importance of psychological and behavioural factors in asset pricing. Studies done by behavioural finance economists have largely shown that stock prices are predictable and that it is possible to earn a riskless profit based on historic prices as well as the use of certain fundamental valuation metrics (Malkiel, 2003). While classical finance postulates that investors are *rational* and build their portfolios using mean-variance optimisation, behavioural finance suggests that investors are *normal* individuals who build their portfolios using the behavioural portfolio theory (Statman, 2019). Behavioural finance

proponents have challenged the theoretical underpinnings of the EMH as well as the accompanying empirical evidence by arguing that investors are not perfectly rational and therefore capital markets are not perfectly efficient. By trying to give a more plausible explanation of asset pricing, behavioural finance incorporates concepts from the diverse fields of finance, psychology and sociology (Shiller, 2003).

A single feature of behavioural finance that has preoccupied researchers in recent times is the concept of investor sentiment. Peterson (2016) argues that a universal definition of sentiment in the finance literature does not exist. Some definitions found in the literature range from the general investor attitude towards capital markets (Shleifer & Summers, 1990); specific beliefs that culminate in either undereating or overreacting to security returns (Barberis, Shleifer & Vishny, 1998) to the optimism of investors towards financial markets (Baker & Wurgler, 2006). Traditional finance models disparage the efficacy of sentiment in explaining asset prices by arguing that suboptimal trading decisions caused by cognitive biases and misguided beliefs will ultimately be eliminated by arbitrageurs.

Black (1986) suggests that irrational investors, also called noise traders, are known for not trading on fundamental information but are instead driven by sentiment. Noise traders are assumed to be individuals who often become irrational and unpredictable. This arises from processing information with finite cognitive capacity and this ultimately leads to volatility as prices deviate from their equilibrium prices. Noise traders will, therefore, be able to make superior returns since limits to arbitrage prevent a perfect adjustment of mispricings (Baker & Wurgler, 2006). The power of noise traders to move prices can be seen in an incident that took place in 2008 when the 2002 news about the bankruptcy of United Airlines resurfaced (Carvalho, Klagge & Moench, 2011). The value of the company fell by 76% in a matter of seconds because noise traders believed the news to be new. Though the subsequent authentication of the news as old led to a rebound of the stock, it remained 11.2% short of its opening price by the end of the day.

To comprehend the deviations from market efficiency, behavioural models seek to establish the nature of irrationality behind investor behaviour. This approach has led to the utilisation of various

biases in behavioural finance theory. Barberis and Thaler (2003) classify these deviations from rational behaviour into biases in beliefs or biases in preferences. These biases are explored below:

- ***Biases in beliefs***

Huberman and Regev (2001) indicate that the continued use of sentiment indices to predict the direction of the market is premised on the confirmation errors of investors who focus on confirming cases while overlooking disconfirming errors. Ultimately, confirmation errors in capital markets culminate in a class of investor who expects higher returns, participates frequently in financial markets but with lower realised returns. This is confirmed by Statman (2019) in his book *"Finance for normal people"*, who posits that investors who trust in their ability to pick winner stocks are frequently oblivious of their deficiencies when they pick loser stocks. These investors, by recognising gains as confirming their abilities in picking stocks and abstaining from recognising losses, never accept losses as disconfirming evidence. Statman (2019) demonstrates the role of confirmation errors in online message boards by arguing that investors tend to commit confirmation errors when they process the information on message boards but assign less weight to disconfirming opinions.

Another strand of literature explores beliefs emanating from representativeness and conservativeness. According to experimental psychologists, there is a tendency for individuals to depend too much on small samples by viewing them as overly representative of the underlying population (Kahneman & Tversky, 1979). Economic agents have also been found to rely too little on large samples to update their priors conservatively. This line of thought led Barberis, Shleifer and Vishny (1998) to model investor sentiment in a context where investors overweight new information relative to previous (representativeness) while at other times underweighting new information (conservativeness). The advent of stock microblogs has given noise traders access to lots of information and these investors may be influenced by these biases in the course of picking their stocks.

Another of the most cited cognitive biases in literature is the overconfidence bias. Daniel, Hirshleifer and Subrahmanyam (1998) define overconfidence as existing when an economic agent

overweights private information compared to public information signals. This leads to a typical investor who depends too much on personal judgement and self-belief in making trading decisions. According to Alpert and Raiffa (1982), overconfidence in the financial markets often manifests itself when investors use too narrow confidence intervals in estimates. The consequence of overconfidence in financial markets is that investors will tend to overlook factual and statistical information in preference for anecdotal information. Daniel, Hirshleifer and Subrahmanyam (1998) report that overconfidence manifests when investors overreact to private information signals. García (2013) demonstrates that investors can view messages from those they follow on stock microblogs like Twitter as private information. Kahneman (2003) argues that overconfidence is especially rife in an environment with low information and demonstrates that overconfidence is independent of the level of expertise in the subject matter.

- ***Biases in preferences***

Besides the biases in belief outlined in the previous section, the literature also explains deviations from rationality using biases in preferences. Most of the explanations of deviations from rationality utilising preference-based deviations are based on the Prospect Theory of Kahneman and Tversky (1979). In Prospect Theory, the utility function is presumed to be convex in the region of losses, kinked at zero and concave in the region of gains. Assuming that X_{t+1} represents the gain or loss an economic agent realises between time t and $t+1$, the utility function v follows the following form:

$$v(X_{t+1}) = \begin{cases} X_{t+1} & for \quad X_{t+1} \geq 0 \\ \lambda X_{t+1} & \quad\quad\; X_{t+1} < 0 \end{cases} \qquad \text{Equation [1]}$$

The cornerstone of Prospect Theory, loss aversion, is captured by the parameter λ, in Equation [1] above. Loss aversion indicates the propensity for individuals to be more sensitive to reductions in wealth than increases. The theory provides the basis on which various biases can be derived relative to an economic agent whose behaviour is in line with the von Neumann-Morgenstern axioms. Prospect Theory postulates that investors are likely to exhibit risk-seeking behaviour in the region of losses. Kahneman and Tversky (1979, p.287) summarise this risk-seeking behaviour from the following citation, "*a person who has not made peace with his losses is likely to accept gambles that would be unacceptable to him otherwise*". This means that applied to financial

markets, the magnitude of investor sentiment is likely to be asymmetric between bull and bear markets.

While Prospect Theory leads to risk-seeking behaviour in the region of losses as explained above, Thaler and Johnson (1990) depart from this phenomenon by suggesting that the risk tolerance of economic agents increase as their wealth exceeds the reference point. Thaler and Johnson (1990) call this the *'house money effect'*. Barberis *et al*., (2001) use this phenomenon to model investors as accepting more risk when they get returns from a portfolio of risky assets that exceed some historical benchmark. The preference-based biases explained above explain some of the reasons why investor sentiment significantly predicts returns in the extreme states of the markets as shown in the survey of literature in Section 2.6.

2.3 Reconciling the EMH and behavioural finance

While finance discourse has been dominated by advocates of the EMH and behavioural finance paradigms, there is an emerging crop of scholars who do not subscribe to the notion of fully efficient markets, neither to markets that can be explained solely by behavioural theories. Awarding the 2013 Nobel Prize in Economics jointly to Robert Shiller and Eugene Fama (Statman, 2018), respectively considered the fathers of behavioural finance and efficient markets, shows the importance of both ideologies in finance. Lo (2004)'s Adaptive Market Hypothesis (AMH) attempts to reconcile the two ideologies by using explanations from evolutionary sciences. The evolutionary nature of financial markets was first propounded by Farmer and Lo (1999) who contest the stationary EMH by arguing that markets are characterised by the *"competition for survival"* and that this leads to cycles of efficiency and inefficiency. The AMH views asset prices in financial markets as reflecting a combination of environmental factors as well as the nature and number of participants *(species)* in the environment. The diversity of the market participants (such as. naïve investors, smart investors, market makers) makes financial markets' efficiency context-specific.

The AMH postulates that investors take market positions in response to the nature and number of market participants and the magnitude of potentially gainful opportunities. Investment strategies in financial markets, therefore, go through cycles of profit and loss depending on the environment.

The rationale of the AMH is that markets are cyclical and are characterised by intermittent periods of efficiency and inefficiency. This cyclical nature of financial markets means that stakeholders have to constantly acclimatise to market conditions because of the dynamic nature of the risk-reward relationships (Lo, 2004). Several empirical studies have since been done that confirm the validity of the AMH. In a study using a sample period spanning from the 1960s to the 2000s, Ito and Sugiyama (2009) report that market inefficiency in the United States financial markets is time-varying. Coming to emerging markets, Seetharam, Auret and Celik (2017) provide evidence of the cyclical tendency of market efficiency in South Africa by indicating that the JSE is weak-form efficient at daily and weekly frequencies while inefficient using lower return frequencies. This is corroborated by Heymans and Santana (2018) who report that the JSE all-share index is weak-form efficient and that the sub-indices move from periods of efficiency to periods of relative inefficiency.

Pedersen (2015) concurs with the AMH by arguing that financial markets cannot be described by the polar viewpoints insinuated by classical finance and behavioural finance respectively. According to Pedersen (2015), although it is possible to beat the market, it is very hard to do so as there are many individual and institutional factors that interplay to prevent this. He, therefore, proposes that instead of financial markets being strictly efficient or behavioural, they are *"efficiently inefficient"*. This means that financial markets are (i) just inefficient enough that investors get compensated for taking a risk and (ii) just efficient enough that active management strategies do not get adequately compensated to encourage the entrance of new capital and new money managers. This view of financial markets is consistent with the AMH as it demonstrates capital markets as constantly evolving and gravitating towards an *"efficient level of inefficiency"* (Pedersen, 2015: 11).

The validity of the AMH has also been supported by practitioners in the investments industry. The belief that markets are not always efficient and that they can be beaten sometimes has reportedly earned some notable investment industry practitioners significant risk-adjusted returns. A well known American company, Rennaissance Technologies, is famed for consistently beating the market through its Medallion Fund. From 1998 to 2019, the Fund returned 66% gross annual returns and after-fees annual returns of 39% (Stevens, 2019). The company's founder Jim

Simmons, who had experience in academia before he moved into the investment industry, so much believed that markets can be beaten to an extent of deploying an employee base with more than 60 PhD holders in quantum physics, astronomy, artificial intelligence, computational linguistics and mathematics just to exploit mispriced situations (Zuckerman, 2019). The company's Chief Executive Officer, Robert Leroy Mercer emphasised the dynamic nature of market efficiency of markets when he said: *"We're right 50.75 percent of the time . . . but we're 100 percent right 50.75 percent of the time"* (Zuckerman, 2019). This evidence from Rennaissance Technologies demonstrates the feasibility of markets to be beaten and illustrates the possibility of markets to be predicted at times.

2.4 Social media and textual sentiment

Theoretically, one would presume that social media platforms are expected to enhance market efficiency as information should spread more rapidly compared to traditional news platforms. On the other hand, as outlined in the following section, social media may help in creating profit opportunities for investors by encouraging emotional responses to the news and being susceptible to noise trading. Clement (2020) defines social media as *"a group of internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of user-generated content".* Some important social media platforms that are common today include social networking platforms (such as Google+ and Facebook), microblogs (such as Twitter and StockTwits) and content platforms (such as YouTube and TikTok). Clement (2020) estimates that the number of social media users globally has been on an upward trend and is projected to increase from the current levels of 3.8 billion users in 2020 to about 5 billion in 2025. In South Africa, the use of social media by citizens has been on an upward trend, with the trajectory expected into the foreseeable future as shown in Figure 1 below:

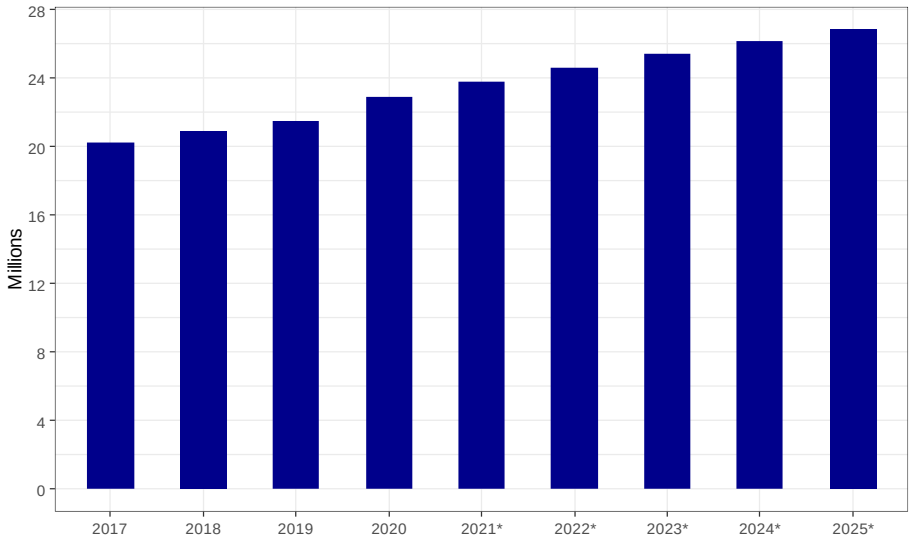


Figure 1: Number of social media users in South Africa according to Statista
Note: ** Indicates estimates for the years 2021-2025*

The notion of textual sentiment from social media is an emerging form of investor sentiment which is text-based and portrays the level of positivity and negativity in texts from social media platforms like blogs and microblogs (Li, van Dalen & van Rees, 2018). This form of investor sentiment has been driven by the proliferation of algorithms that deduct sentiment from text and the parallel increase in the interactions among investors using microblogs, social media sites, discussion forums and internet message boards. Many individual traders are abandoning traditional news platforms for social media platforms as the former allows users to express their opinions in addition to obtaining information (Kearney & Liu, 2014). Twitter and StockTwits have emerged as some of the most used online microblogs that finance and computer science researchers are using to extract textual sentiment (Li, van Dalen & van Rees, 2018). Twitter Inc. was formed in the year 2006 and provides a platform that allows participants to post 280 characters expressing their opinion to which other participants can reply, like or retweet. StockTwits Inc. was founded two years later in 2008 and uses the same format as Twitter but the forum is exclusively for traders and investors. Participants of both Twitter and StockTwits forums can use a dollar sign ($) preceding a ticker symbol to denote that they have been talking about a particular stock (for example, Sasol is represented as $SOL in a *tweet*).

Textual sentiment analysis relies on machine learning – which involves the utilisation of statistical procedures to deduce the details of texts as well as categorising them using statistical inference (Kearney & Liu, 2014). The machine learning process starts with the analysis of a complete set of texts, which constitutes the training set. The words in the training *corpus* are then categorised manually as "positive", "negative" and "neutral" or any other classification dimension (such as "strong", "weak" or "passive"). An analysis algorithm (such as Naïve Bayesian Algorithm) is subsequently selected and deployed on the training corpus. The selected algorithm will learn the instructions for the classification of sentiment from a pre-classified dataset and applies these guidelines out of sample and to the entire training set.

Hirshleifer and Teoh (2009) emphasise the importance of information transmission in financial markets so as to fully comprehend the theory of capital markets. Retail investors in the real world are known for observing the behaviour of peers and in the process gather information either through conversations or *via* social media (Bartov, Faurel & Mohanram, 2018). The information that is transmitted from one individual to another through social networks or otherwise is, unfortunately, not processed through Bayesian updating as insinuated by traditional finance models, but through both reasoning and emotional factors (Twedt & Rees, 2011). The advent and propagation of social media platforms and microblogs have increased the scope and speed of this information transmission among investors.

The developments in FinTech in recent years has led to the influx of retail investors into the stock market and these investors have been known to have an affinity for social media for informational purposes rather than consulting professional money managers (Twedt & Rees, 2011). According to Moss, Naughton and Wang (2020), the emergence of broker-dealer applications like Robinhood in the USA has drastically reduced the costs of trading and in the process attracted an unprecedented number of retail investors. In the USA, the COVID-19 pandemic revolutionalised financial markets as many novice investors entered into financial markets either as a result of the boredom from pandemic-induced lockdowns, increased liquidity from stimulus cheques or speculative tendencies by traditional sporting gamblers who switched to the stock market because of the closure of betting houses (Moss, Naughton & Wang, 2020). These traders have been reported to closely follow other retail investors through social interactions and deciphering how other

investors are investing. The effect of herding in financial markets propagated by social media platforms can be seen in the filing of bankruptcy by an American car hire company Hertz Inc. Soon after filing for bankruptcy, Robinhood investors snatched up the stock leading to a rally of 900%, against what traditional finance theory would predict (Moss, Naughton & Wang, 2020). In South Africa, the launch of a Robinhood-like company, EasyEquities in 2014, has significantly increased the participation of retail investors on the JSE. Thus, it would be interesting to see if sentiment from social media, given the increase in the number of noise traders, can explain some stock market features.

2.5 Social media and wisdom of the crowds

The rising importance of the use of online platforms in the investment fraternity has brought to the fore a very important concept, the *"wisdom of the crowds".* The notion was first formulated by Galton (1907) who found out that on average, a multitude of users led to more informed decisions compared to a few experts or professionals. According to this concept, individual estimates contain subjective errors as individuals are prone to their specific preferences and expectations (Mesgari *et al*., 2015). However, these subjective errors are bound to dissipate when all the individual estimates are aggregated together while at the same time preserving the true information (Mesgari *et al*., 2015).

The Fourth Industrial Revolution and the easy and cheap access to the internet it has brought, has greatly propelled the internet to playing an intermediating role in the *"wisdom of the crowds"* concept. Wikipedia, a website that went live in 2001 and today hosts more than 35 million articles across more than 250 languages of the world, is one way in which the *"wisdom of the crowds"* phenomenon has been witnessed through the internet. According to Mesgari *et al*., (2015), the entries on Wikipedia entered by multitudes of people from different backgrounds and qualifications are of a high quality that compares favourably with the quality from entries written by experts and professionals in encyclopedias. This power of the *"wisdom of the crowds"* is also expected in the area of finance as the swarm intelligence from the multitude of social media users is presumed to help investors predict the stock market. Using a social investment platform called *Sharewise,* Breitmayer, Massari and Pelster (2019) examine whether the *"wisdom of the crowds"*

consensus has explanatory power on the stock market returns. The study finds that the cumulative intelligence of the masses has significant predictive power on returns.

.

2.6 The association between textual sentiment and stock market features

Some studies have explored the nexus between *tweet* features and stock market features as well as the information content of past scores of *tweet* features that could be used to predict future values of stock returns. Tetlock (2007) inspects the causal association between a media pessimism factor and asset returns using vector autoregressive analysis. The media pessimism factor is constructed from a Wall Street Journal column ("Abreast of the Market") utilising the principal component analysis of 77 preprogrammed General Inquiry categories from the Harvard Psychosocial dictionary. A single media pessimism factor is then created from collapsing the 77 categories. Results from the study attest to the prowess of media in explaining asset returns as pessimism is associated with downward changes in prices the next day. Tetlock (2007) also reports high volatility of the Dow Jones Index resulting from a high pessimism factor.

An early study that examined the impact of message boards on the stock market, upon which most subsequent studies are based, is the study by Antweiler and Frank (2004). Using 45 Dow Jones listed companies and more than a million messages mined from Raging Bull and Yahoo! Finance, the authors utilise computational linguistics to examine if a stock message can help to forecast share returns and volatility. The authors conclude that online messages can be used as a strong proxy for noninformational transactions. Antweiler and Frank (2004) also find that investor disagreement is connected with increased trading volume. This corroborates the traditional finance theory which postulates that disagreement among investors is central to trading (Karpoff, 1986).

One of the earliest documented studies to utilise textual sentiment extracted from Twitter was done by Bollen, Mao and Zeng (2011). The study made use of six measures of public mood namely; *"calm", "alert", "sure", "vital", "kind"* and *"happy"*. Granger non-causality and a Self-Organising Fuzzy Neural Network were subsequently employed to test the forecasting power of the above-mentioned public mood states on the Dow Jones closing prices. The study concluded that public mood states can predict the changes in the closing prices of the Dow Jones. In a later study, Lachanski and Pav (2017) disparage the results of the Bollen, Mao and Zeng (2011)

(hereinafter referred to BMZ) study on various grounds. Firstly, they question the decision by BMZ to restrict their study sample to the period 28 February 2008 to 3 November 2008, yet they had access to data covering the entire 2006 to 2008 period. Secondly, Lachanski and Pav (2017) argue that BMZ failed to position their findings in the context of the equilibrium theories of investor sentiment as seen for example when they relate their findings to earlier work done on the impact of *tweet* sentiment on box office receipts, a reference not related to any equilibrium theories of investor sentiment. Lachanski and Pav (2017) also noted some methodological irregularities in the BMZ paper and, after replicating the BMZ study, concluded that there was no evidence that *tweet* features can predict the Dow Jones, at least within the sample period used.

Sprenger *et al*., (2014) expand on the previous study done by Antweiler and Frank (2004) by only analysing specific stock microblogging messages compared to the previous approaches which used all available Twitter messages. More than 250 000 stock-related *tweets* are used to examine if sentiment from *tweets* can impact stock returns, volatility and trading volume. Replicating Antweiler and Frank (2004), the authors use fixed-effects OLS regressions to test the contemporaneous relationship between *tweet* features and stock market features and Fama and Macbeth style-regressions to examine the lead-lag relationships. The conclusion from the study shows that stock-related *tweets* have information content that has not yet been assimilated by the market. The conclusion is based on the findings which show that bullishness, message volume and message disagreement are positively and significantly related to the stock market features of stock returns, volatility and trading volume.

Li, van Dalen and van Rees (2018) replicate the methodology used by Antweiler and Frank (2004) but using stock-related *tweets* instead of stock-related news. The authors assemble more than one million *tweets* for 100 companies listed on the S&P500. Computational linguistics is used to detect spam, remove slang as well as to detect language from the *tweets* which were gathered via the Twitter Application Programming Interface (API). The other difference between the approach used by Li, van Dalen and van Rees (2018) compared to Antweiler and Frank (2004) is the granularity of the analysis. Whereas the latter used daily granularity in their sentiment analysis, the former used 15-minute-interval intraday granularity which is more relevant for real-time microblogs like Twitter. The naïve Bayesian method was used on a training set of 2144 messages

and these were manually classified into three signals, namely; buy, hold and sell signals. The study examined the impact of two message features; namely, message volume and level of agreement on stock features. The findings from the study show that the message features used had a positive effect on stock returns, trading volume and volatility. Consistent with Antweiler and Frank (2004), Li, van Dalen and van Rees (2018) report that disagreement induces trading.

The literature shown above clearly shows that the mainstream studies done on the link between sentiment mined from social media and the stock market are concentrated in the developed markets of America and the United Kingdom. This is in line with Indaco (2018) who posits that research has established that the majority of social media users are clustered in developed countries. Given the increasing importance of the emerging markets within the global financial markets and the increasing internet penetration, it would be of paramount importance to understand whether social media influences the financial markets in these emerging markets.

Because western social media platforms are banned in China, Xu *et al*. (2017) use an indigenous Chinese Twitter-like social media platform, *Sina Weibo*, to extract textual sentiment. Machine learning is used to extract sentiment from Weibo microblogs and this sentiment is categorised into *"sadness", "fear", "disgust", "anger"* and *"joy"*. Out of the four metrics used to measure sentiment, sadness is found to have a more pronounced contemporaneous association with stock returns. On the lead-lag relationships, stock returns are found to cause Weibo sentiments than the other way round. Message disagreement is also found to be negatively associated with trading volume. This is consistent with the "no-trade theorem" which states that disagreement induces no trading as it necessitates the review of prices and opinions (Milgrom & Stokey, 1982).

Steyn *et al*., (2020) depart from previous studies which concentrate on developed financial markets by including both developed and emerging countries in a cross-national study. They utilise sentiment extracted from Twitter microblogs using three different machine learning classification algorithms. The peculiarity of the study, according to the authors, is in the utilisation of the '*syuzhet*' package which more accurately captures sector-specific sentiment terms compared to the conventional lexicons that are usually used by researchers like Wordnet and WordNet-Affect. The study uses data from the developed economies of the United Kingdom, Spain, France, Germany,

the USA and the emerging economies of India and Poland. The emotion scores that are used in the study include *"anger", "fear", "anticipation"* and *"trust".* Using Variable Importance Analysis (VIA), the study reports that sentiment extracted from *tweets* is as significant in predicting the stock market returns in developed countries as in developing countries.

Daszynska-Zygadlo, Szpulak and Szyszka (2014) include a larger sample of emerging markets in their study that sought to confirm the contemporaneous association between optimism extracted from news and the stock market. The study utilised data from eight emerging economies namely; China, Brazil, India, Turkey, Mexico, Russia, Poland and South Africa. Investor sentiment is extracted from the Thomas Reuters MarketPsych Indices which use machine learning and big data techniques to extract sentiment from media coverage. The aggregate database ultimately consisted of more than 2 million real-time news articles collected from more than 50 000 premium news providers between 2003 and 2013. Where most previous studies used daily and intraday granularity in media sentiment analysis, this study departed from this norm and opted for weekly granularity. The authors used factorial regression to establish the nexus between their media-extracted sentiment index and excess returns. The findings from the study show that in Poland, South Africa, Russia and Turkey, investor sentiment is not significantly related to excess returns while a positive and significant association is reported for China, Brazil, Mexico and Poland. For the countries which reported a positive and significant link between sentiment from news content and excess returns, the study further establishes that excess returns are more pronounced during periods of negative sentiment.

In South Africa, various studies have been done on how investor sentiment affects the stock market. Studies done have attempted to establish the link between investor sentiment and a host of dependent variables like stock market returns (Dalika & Seetharam, 2015); return volatility (Rupande, Muguto & Muzindutsi, 2019) and foreign financial flows (Muguto, Rupande & Muzindutsi, 2019). These studies have mainly used survey-based proxies for investor sentiment. The only known study that utilised sentiment extracted from Twitter in South Africa at the time of this present study is the study by Johnston and Maree (2015). The study by Johnston and Maree (2015) sought to establish if moods from stock-related *tweets* can significantly predict the JSE All Share Price Index. More than 3.9 million *tweets* spread over 55 days were gathered using Twitter's

API. Instead of using the popular Profile of Mood States (POMS) to come up with mood states, the authors elected to use the Extended Profile of Mood States (XPOMS) with six mood states namely *"anger", "vigour", "confusion", "fatigue", "depression"* and *"tension"*. The study reported significant Granger causality from fatigue to JALSH. The present study, therefore, builds on the work of Johnston and Maree (2015) by covering a fairly long period of 5 years and using firm-specific *tweet* features compared to market-specific *tweet* features used by Johnston and Maree (2015).

2.7 The asymmetric effects of sentiment on returns

Although various studies have been done to establish the association between sentiment from microblogs as seen in Section 2.5 above, there is still widespread debate on when textual sentiment matters most for investors. Studies that have been done to investigate when investor sentiment matters more have largely produced mixed results. While Shleifer and Vishny (1997) show that the impact of sentiment on share returns should be symmetric, a complex relationship is likely to arise because of the variations in shorting expenses from different market conditions. Stambaugh, Yu and Yuan (2012) posit that the differences in shorting costs and therefore shorting impediments in different market conditions should make overpricing prevail over underpricing. *Smart investors* are therefore more probable to enter capital markets following bullish rather than bearish market conditions and hence predictability should be more pronounced during good market conditions (Lemmon & Portniaguina, 2006).

Several empirical studies have been done which confirm the above-mentioned phenomenon; that the predictability of stock returns is more pronounced under tranquil market conditions. You, Guo and Peng (2017) investigate whether investor sentiment has forecasting power on returns across some 10 international equities markets by utilising a novel sentiment proxy mined from Twitter. The authors use a quantile-based ($\tau \in [0.05, 0.95]$) Granger non-causality test. The empirical results from the study show that the causal associations between a *tweet* happiness index and stock returns vary across the range of the quantiles used. The causal effect of investor sentiment on stock returns is found to be statistically significant only in high return quantile intervals. On the other hand, the causal effect of returns on investor sentiment exists in the tail quantile of returns only.

Thus, the results from the study demonstrate the weakness of sentiment in influencing stock returns during bad market conditions but significantly affecting returns in good market conditions.

Chakraborty and Subramaniam (2020) create a market-based metric for investor sentiment (Market Mood Index) to examine the asymmetric association between sentiment and stock returns using a sample of companies on the Indian Stock Exchange. The authors use the quantile causality approach to investigate this relationship with the results showing that investor sentiment has predictive power only in extreme return quantiles. A positive and significant causal effect of investor sentiment on stock returns is reported for lower return quantiles ($\tau \in [0.05 - 0.2]$), while at higher quantiles ($\tau \in [0.8 - 0.95]$) high investor sentiment is found to negatively predict stock returns.

Contrary to the above-mentioned phenomenon, Prospect Theory suggests that the expected utility theorem does not adequately capture human behaviour in the face of gains or losses (Kahneman & Tversky, 1979). According to Prospect Theory, individual investors are susceptible to significant pain from a loss compared to excitement from a gain of comparable magnitude (Kahneman & Tversky, 1979). This proposition asserts that the behaviour of investors varies significantly, contingent upon the state of the capital markets and whether they are characterised by periods of anxiety and fear or by prosperity and tranquillity. Related to this, studies have also shown that investors are more likely to be irrational during periods of anxiety (Gino, Brooks & Schweitzer, 2012). All this points to a more pronounced prediction of stock market features during bad times compared to good times. Ma, Xiao and Ma (2018) report results that are consistent with this phenomenon, using sentiment extracted from macroeconomic variables through principal component analysis. The study uses a quantile regression approach and concludes that the prediction of stock returns using investor sentiment is more pronounced at lower return quantiles ($\tau \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6,\ 0.7]$). However, at upper quantiles ($\tau \in [0.8, 0.9]$), the forecasting power of sentiment is lost and the regression coefficients become anomalous.

The asymmetric effect of investor sentiment on stock returns is also reported by Swamy and Dharani (2019). The study uses a media-based proxy for investor sentiment, google search

intensity. Data from Indian Stock Exchange-listed companies spanning 5 years from 2012 to 2017 are analysed with results showing the predictive ability of google stock searches on stock returns. The quantile regression results obtained also show that the forecasting ability of investor sentiment is more pronounced at lower quantiles ($\tau \in [0.05, 0.25, 0.5]$) of returns with higher quantile ($\tau \in [0.75, 0.95]$) returns having no significant predictive power.

Using a range of event studies and social media sentiment mined from Twitter and StockTwits, Agrawal *et al*., (2018) report that using intraday granularity, social media sentiment has significant effects on stock turnover and returns. Particularly, negative social media sentiment is seen having a more pronounced and significant effect on turnover and returns compared to optimism from social media. Kim and Kim (2014) analyse a dataset of more than 32 million messages posted on Yahoo! Finance for 91 companies to examine the behaviour of investor sentiment at the tail ends of stock return distributions. To do this, the authors describe an extreme change in the stock price as one that changes by either below or above two standard deviations from the expected returns of a stock. The previous 240 days are used to calculate the average returns as well as the standard deviations of the returns. The findings from the study show that sentiment can explain stock returns significantly at lower than two standard deviations from normal returns.

Chau, Deesomsak and Koutmos (2016) investigate the importance of sentiment in the US capital markets and how investor sentiment influences the trading behaviour of investors. The authors particularly utilise market-based and survey-based sentiment indices in their study. The findings from the study firstly establish that sentiment-based traders are not irrational investors as presumed by classical finance theories as they also trade against the crowd. The study also suggests that sentiment-driven traders can also serve as *smart investors* by providing liquidity to the market by buying overly sold stocks. A notable result from the study confirms that sentiment-driven trading increases and is more aggressive during bear conditions than during bullish markets.

Another strand of literature has also strived to empirically test whether investor disagreement matters more in bear markets or bull markets. Li and Wu (2014) examine the relationship between analyst forecast dispersion using a quantile regression approach. The authors use forecast diffusion

among analysts as a proxy for the level of investor agreement. A sample of 4122 non-financial companies registered in the United States of America is used in an analysis that utilises observations spanning the period 1983 to 2009. The empirical findings from the study indicate that analyst disagreement as measured by the dispersion of opinion index is dynamic and oscillates across the different return quantiles. Specifically, the study reports that the association between analyst disagreement and stock returns is significantly negative (positive) for return quantiles between 0.05 and 0.60 (between 0.80 and 0.95). On the other hand, a return quantile between 0.65 and 0.75 is associated with analyst opinion dispersion which is insignificantly different from zero.

While Li and Wu (2014) use analyst forecast deviation as a representation of investor disagreement, Hillert, Jacobs and Müller (2018) utilise disagreement among journalists as a proxy for investor disagreement. The authors analyse the texts from a novel database of 729 000 media articles extracted from major news outlets between 1989 and 2010. The results from the study show that the disagreement metric constructed to measure journalist disagreement is inversely associated with market returns and that the connection is more pronounced during bear periods.

The results reported by Li and Wu (2014) and Hillert, Jacobs and Müller (2018) are in line with Diether, Malloy and Scherbina (2002)'s proposition that traders tend to overprice shares with high levels of investor opinion dispersion. Subsequently, the correction process of the overpricing leads to variations in the response of stock returns to analyst opinion dispersion at different return quantiles. Diether, Malloy and Scherbina (2002) propose that when the overpricing is corrected, stock returns analyst opinion differences should be inversely related to returns, while a direct connection between analyst opinion dispersion and stock returns should exist when the overpricing is continuing. Ultimately, when the overpricing correction process is complete, there should be a trivial nexus between investor disagreement and stock returns.

While the above studies show the relationship between investor disagreement and stock returns at diverse return quantiles, the proxies for investor disagreement (journalist disagreement and analyst opinion dispersion) can best be described as expert opinions and might not reflect the nature of disagreements that are exhibited on stock microblogs like Twitter and StockTwits. Al-Nasseri and Menla Ali (2018) bring another dimension to the effect of investor disagreement by utilising a

proxy for agreement constructed from 289 443 online *tweets* from StockTwits. The authors specifically examine potential asymmetric disagreement effects on stock returns in bear and bull markets. The study adopted the Antweiler and Frank (2004) agreement index as well as an agreement index calculated from the variance of the bullishness index as a robustness check. The top 30 highly traded stocks of the Dow Jones are used in the analysis. Traditional ordinary least squares regression permitting for firm-fixed effects is used to examine how disagreement fluctuates across various states of the market. The study reports that stock returns negatively (positively) respond to investor disagreement during bull (bear) periods. These results demonstrate that investor opinion dispersion can be viewed as a proxy for risk in bear periods.

2.8 Methods and models used for textual sentiment examination in finance

This section looks at the econometric models that researchers have mostly used to comprehend the role of textual sentiment in capital markets. Studies reviewed in the previous sections largely show mixed results and a survey of the methods used may help to explain some of these diverging findings. This section, therefore, looks at the different sources of information used in textual analysis, the different methods used in content analysis, the different metrics used for measuring textual sentiment and the econometric models used for examining the association between textual sentiment and various stock market features.

2.8.1 Sources of information

According to Kearney and Liu (2014), the sources of the qualitative information that have been used in textual sentiment studies in finance can be categorised into three broad categories, namely public filings, news articles and messages from the internet. Kearney and Liu (2014) posit that the regulatory disclosures made by public companies present an avenue through which researchers can extract textual sentiment. Particularly, the style of the language used in these public disclosures, including the tone, can be a useful indicator of the expected future performance of a company outside the information contained in annual reports. On the downside, these public corporate disclosures are usually released at low-frequency intervals either quarterly, semi-annually or annually, thereby reducing the efficacy of the information in a contemporary fast-paced corporate environment. Thus, studies that use textual sentiment extracted from public filings

are few and an example is Loughran and Mcdonald (2011) who use textual sentiment from 10-K filings[4].

An increasing source of content for textual analysis in research is the sentiment that is extracted from media sources. The media-expressed textual sentiment is the sentiment that is extracted from news media, analyst briefings or commentaries about the performance of companies. The advantage of this source of information over the corporate filings discussed above is that media-expressed sentiment can be obtained at higher daily and intraday intervals. Various studies have utilised textual sentiment extracted from the media as representations of investor sentiment. Tetlock (2007) extracts textual sentiment from investment and finance news using a major American newspaper, the Wall Street Journal. In a follow-up study, Garcia (2012) uses a media-expressed sentiment from news articles canvassed from another top American financial news medium, The New York Times. Both studies used daily and intraday frequencies which demonstrates the advantage of this source of textual information compared to public filings.

Internet-expressed textual sentiment has become more important in investor sentiment studies nowadays on the back of the increase in the number of user-generated content websites. Most people nowadays spend considerable time on the internet reading, posting and reacting to opinions from peers. Kearney and Liu (2014) however, note that internet-expressed sentiment is mostly noisier than the other sources of information for textual sentiment outlined above. This is mainly because internet-based textual sentiment is driven by individuals who are not regulated compared to corporates and media houses, who are regulated and are usually guided by specific codes of conduct when they post information online. Internet-based computational textual analysis in finance has been done using information from various sources such as Yahoo! Finance (Antweiler & Frank, 2004), Yahoo message boards (Das, Martínez-Jerez & Tufano, 2005), and Twitter and StockTwits (You, Guo & Peng, 2017; Steyn *et al*., 2020). The major strength of online-based sentiment is that the information is captured in real-time and the influence on the stocks can be matched in real-time. The present study utilises the internet-expressed sentiment extracted from

[4] Statutory compliance reports required by the Securities and Exchange Commission (SEC) for all public companies in the USA and show a summary of a company's performance. 10-K filings typically contain more information compared to a company's annual report.

Twitter and StockTwits by Bloomberg Inc. and the computational specifications are outlined in Section 3.2.1.

2.8.2 Content analysis methods

According to Kearney and Liu (2014), the content that is collected from texts in the finance literature is commonly analysed using dictionary-based as well as approaches based on machine learning (ML). These approaches are discussed in this section.

2.8.2.1 Dictionary-based approaches

These approaches utilise algorithms that necessitate the reading of texts and the subsequent categorisation of the words or phrases so extracted using a dictionary that is defined before the analysis. The General Inquiry (GI) has been used by various studies that extract sentiment using dictionary-based approaches including Tetlock (2007) who uses content from the Wall Street Journal and Twedt & Rees (2011) who use content from companies' regulatory filings. *DICTION* is another popular source of word lists that have gained momentum in finance studies (Kearney & Liu, 2014). Though *DICTION* was founded by someone who is an expert in politics and mass media, the dictionary is based on 33 separate dictionaries and this has seen it being widely used in finance because of its scope (such as Loughran & Mcdonald, 2011)

The main disadvantage of the two sources of word lists mentioned above is that they are based on the English language and may not capture sector-specific terms and technical words used in the broad area of finance. An example of this anomaly is the categorisation of words like "tax" and "liability" as capturing negative sentiment from the GI/Harvard dictionary, yet these words do not reflect negativity in the financial context (Loughran & Mcdonald, 2011). Loughran and Mcdonald (2011) substantiate this by stating that more than three out of every four words which are categorised as negative words using the GI/Harvard lexicon are not negative words in the economic context. This has led researchers to formulate dictionaries/lexicons which are relevant to the finance field to capture sentiment expressed from texts accurately. The LM dictionary (Loughran & Mcdonald, 2011) is one such finance-specific dictionary that has been widely used in more recent studies (such as Jegadeesh & Wu, 2013; Huang *et al*., 2014).

Another issue that is important in dictionary-based approaches to content analysis is the weighting of the word lists. Proportional weighting is used in studies that presume equality in the importance of the words in the dictionary. Term weighting, on the other hand, assumes that different words capture different aspects of mood and cannot be equally weighted. In this regard, Jegadeesh and Wu (2013) argue that it is of paramount importance to give a weighting to words relative to the historical market reaction to those words. Jegadeesh and Wu (2013) emphasise the role of weighting in dictionary-based approaches to content analysis by proposing that a more accurate weighting process is desirable compared to a more accurate and complete dictionary.

2.8.2.2 Approaches based on machine learning

ML methods to content analysis have been driven by the rising power of computers as well as the deployment of more sophisticated algorithms in finance. Statisticians and computer programmers have been the main drivers of this approach which classifies texts from documents by utilising statistical inference. Using this approach, a "training set" is formulated based on the complete set of text to be analysed. The training set is then used to come up with a manual classification of the words into predefined categories like *"positive", "negative" or "strong"*, *"weak"* or any other agreed classification criteria. After manual classification of the words from the training set, an algorithm is then trained on the training set. The algorithm chosen (such as the Naïve Bayesian) will *"acquire knowledge"* rules for classifying sentiment extracted from the texts from the training set, and these rules are then applied out of sample to the whole sample of texts.

Kearney and Liu (2014) indicate that the main advantage of dictionary-based approaches over machine learning approaches to content analysis is the fact that the latter is time-consuming and laborious as it needs a manual categorisation of the training set first before the rules can be deployed out of sample. Besides the disadvantages of machine learning-based approaches to textual content analysis cited above, the approaches have been reported to be more accurate compared to dictionary-based approaches. Huang *et al*., (2014) for example, report an accuracy rate of 76.9% in their out of sample machine learning approach compared to the 48.4% and 54.9% obtained from the dictionary-based *GI* and *DICTION* lexicons respectively. It is for this accuracy reason that this study utilises the textual sentiment metrics extracted from Bloomberg Inc. using

supervised machine learning approaches to content analysis which are explained in detail in Section 3.2.1.

2.8.3 Measuring textual sentiment

Kearney and Liu (2014) posit that the extraction of sentiment from texts is the most difficult part of sentiment analysis in finance. After successfully extracting sentiment, measuring it, therefore, becomes a straightforward exercise and the methodology chosen depends on the method of content analysis that would have been adopted. A survey of literature on the measurement of textual sentiment in the finance domain shows that dictionary-based approaches measure sentiment as a proportion of the number of words in a given class of sentiment to the aggregate number of words in the text (Kearney & Liu, 2014). Tetlock *et al*., (2007) use a standardisation percentage *(z-score),* which is especially utilised when the total number of words is not stationary. Twedt and Rees (2011) use a relative measure of sentiment that divides the difference between positive and negative words by the total number of words. Using this metric, a positive figure denotes positive sentiment while a negative figure denotes negative sentiment. Machine learning-based measures of sentiment, on the other hand, give a metric that is based on the classification of texts. For example, a metric of 1 is given if the deployed algorithm forecasts a sentence to be positive and, 0 and -1 if it predicts the sentence to be either neutral or negative respectively.

2.8.4 Econometric models

Kearney and Liu (2014) note that sentiment analysis from texts is a fairly new line of research in finance and that this has seen some novel and innovative methods being utilised for hypotheses testing. According to Kearney and Liu (2014), contemporaneous linear regression models remain the most used methods in textual analysis studies and these mainly utilise market-wide textual sentiment compared to firm-specific textual sentiment. Textual sentiment finance literature is dominated by studies that use some firm performance variable like stock return, volatility and future earnings as the regressand while the main predictor variable is the sentiment index constructed using sources and analyses explained in Sections 2.7.2 and 2.7.3. Various control variables have also been used in these studies, including return momentum, market returns as well as the Small Minus Big factor (Kearney & Liu, 2014). While one of the main assumptions of linear regression models is linearity, some studies have opted for non-linear models. To this end, other

studies have used logistic and probit regressions to model non-linear relationships. Since the regressand should be binary, the idea of logistic regressions is to investigate if sentiment can be significantly explained by a particular event.

Though literature shows that most studies on textual sentiment utilise aggregate market sentiment and aggregate market features, some studies also utilise firm-level sentiment and firm-level market features. Most of the studies that use panel data in textual analysis (such as Twedt & Rees, 2011; Sprenger *et al*., 2014; Agrawal *et al*., 2018) have invariably and consistently used OLS estimations with firm-fixed effects to establish the contemporaneous nexus between textual features and stock market features. Other studies (such as Antweiler & Frank, 2004; Kim & Kim, 2014; Agrawal *et al*., 2018) have also included random effects estimations to ensure the robustness of results.

Some studies use event study methodologies where the linear regression model explained above is used but the response variable is the Cumulative Abnormal Return (CAR) over some determined event window. The various control variables that are utilised in studies include the book-to-market ratio, value of stocks, leverage ratio, momentum and return volatility (Kearney & Liu, 2014). Ordinarily, because event studies seek to investigate the impact of a specific event on firm performance, the studies that use the methodology mostly use corporate filings and media-expressed sentiment compared to internet-based sentiment.

One of the most important features of online messages that researchers have been grappling with in recent times is whether these online messages contain useful information that can be used to predict future values of market features. Methodologically, most studies have used OLS with lagged values of the online features. To find the predictive power of sentiment on returns several studies have also used lagged Fama Macbeth-style regressions (Loughran & Mcdonald, 2011). Li, van Dalen and van Rees, (2018) use Structural Vector Autoregression analysis and Bollen, Mao and Zeng (2011) used Granger non-causality tests to determine the predictive power of investor sentiment on stock returns.

Of late, many researchers have been preoccupied with establishing whether textual sentiment affects the stock market features symmetrically or asymmetrically (such as Li, Shen & Zhang,

2019; Chakraborty & Subramaniam, 2020). While models based on the OLS regression model the mean of the regressand on the mean of the predictor variable, the presumption is that the effect of sentiment on market features is symmetrical. Quantile regression, which seeks to model the effect of the independent variable across the whole distribution of the dependent variable has become particularly important in textual analysis in finance. Quantile regression (explained in detail in Section 3.3.2) has been reported as a robust model estimator for sentiment-based studies as it accommodates many of the violations of OLS like the presence of outliers, heteroscedasticity and nonlinearity. Various studies have since utilised quantile regressions across developed and developing countries (such as You, Guo & Peng, 2017; Chakraborty & Subramaniam, 2020).

2.9 Summary

This section outlined the findings from existing research on the extent to which the stock market can be explained by textual sentiment features. The theoretical underpinnings of the study were explored, starting with the EMH, which hypothesises that markets are informationally efficient such that all past and current information is instantaneously fused into the asset prices. The emergence of the behavioural finance paradigm, which offers potential explanations to a host of anomalies, maniacs and stock crashes which the EMH has failed to explain, was also outlined. According to proponents of behavioural finance, several cognitive behaviours of investors lead to informationally inefficient markets, thus creating opportunities for investors to trade on information. Also outlined in this section is the emergence of textual sentiment as a proxy for investor sentiment and a systematic review of the literature that outlines the role of textual sentiment in financial markets.

Empirical studies on the role of textual sentiment in the stock market were also systematically reviewed. Studies done in developed financial markets and emerging markets largely show mixed results. These differences in the findings could be a result of methodological differences as a survey of the econometric models used in investor sentiment studies shows a variety of methods used. On the link between the magnitude of investor sentiment and stock returns, the majority of the studies confirm the asymmetric effects of sentiment on stock returns. Studies reviewed have largely shown that the effect of investor sentiment on stock returns is more pronounced during bear conditions than during bull and normal market conditions. This is consistent with Prospect

Theory, which argues that economic agents are loss averse and are likely to embark on a risk-taking behaviour after suffering a loss than after getting a market gain of the same magnitude. The next section looks at the methodology that was used to test the hypotheses of the study.

3 DATA AND RESEARCH METHODOLOGY

3.1 Data

The study examines whether firm-level *tweet* features can explain stock market features. To investigate the aforementioned, the study utilises firm-level data for all FTSE/JSE All Share Index (JALSH) constituent firms for the period 1 January 2015 to 31 December 2019. Firm-level daily data on *tweet* features, closing prices, low prices, high prices, trading volume and market capitalisation are extracted from the Bloomberg terminal. Returns are adjusted for corporate actions and/or dividends where applicable. Only the current JALSH constituent companies are included in the study as they are the only companies for which *tweet* features data are available on the Bloomberg terminal. This means companies that were part of the JALSH during the sample period but exited the index before 31 December 2019 are excluded from the analysis.

The sample period is limited to the period after 1 January 2015 as Bloomberg only started incorporating *tweet* features data for the JALSH firms on its platforms in 2015. Since only daily *tweet* features data are available from Bloomberg, the study uses daily granularity analysis, though an intraday granularity would have been more appropriate because of the real-time nature of *tweet* messages. Bloomberg aggregates the *tweets* of a specific listed company at the end of the day where aggregation leads to aggregated *tweets* containing positive, negative and neutral sentiment. This means that the *tweets* of a particular share are aggregated at the end of the day to create a single observation for that day.

3.2 Variables

This section defines the variables used in this study as well as justification for their inclusion in the study:

3.2.1 *Tweet* Features

Following Sprenger (2014), three *tweet* features are used as the explanatory variables: namely; bullishness (B_t), message volume (M_t) and message agreement (A_t). The process of calculating the bullishness index used by Bloomberg Inc. starts with manually analysing large datasets of

tweets using human experts. Labels are then assigned to each *tweet* and categorised into positive, negative and neutral labels using the following question;

> "*if an investor having a long position in the security mentioned were to read this tweet, would he/she be bullish, bearish or neutral on his/her holdings"*

The manually classified feeds are then fed into machine learning models that are taught to imitate language experts in analysing text messages. The completed machine learning models are subsequently used to scrutinise new *tweets* tagged with tickers and assigns each *tweet* a story-level sentiment score ranging from -1 to +1 in real-time. Bloomberg does not, however, disclose the details of the models used to determine the sentiment scores because of their proprietary nature. The average firm-level daily sentiment (bullishness) is then extracted from the weighted average story-level sentiment scores in the last 24 hours collected from Twitter and StockTwits and updated every day 10 minutes before the JSE opens and is calculated as:

$$B_{i,t} = \frac{\sum k \in P(i,T) S_i^k C_i^k}{N_{i,T}}, \qquad T \in [t-24h, t] \qquad \text{Equation [2]}$$

Where:

$B_{i,t}$ is the bullishness score for firm i at time t;

S_i^k is the sentiment polarity score for *tweet* k that references firm i;

C_i^k is the confidence of *tweet* k that references firm i;

$P(i,T)$ is the set of all non-neutral *tweet* feeds that reference firm i in the 24 hour-period, T;

$N_{i,T}$ is firm $i's$ total number of positive or negative *tweets* during period *T*.

Bullishness ranges from -1, the most negative sentiment to +1, the most positive sentiment. This means that a bullishness score of 0 denotes neutral sentiment. Sprenger *et al*., (2014) suggest that the bullishness score accurately captures the quality of a *tweet* about a company and can therefore be used as a proxy for investor sentiment. Several studies have used the bullishness score as a proxy for investor sentiment, especially those studies using textual features extracted from Twitter

and StockTwits (such as Sprenger *et al*., 2014). The bullishness index is therefore used in this study as a proxy for investor sentiment.

Bloomberg provides the daily total number of *tweets* for each firm i aggregated at the end of the day. Message volume ($M_{i,t}$) in this study is calculated as the natural logarithm [$\ln(1 + aggregate\ tweets)$] of the aggregate *tweets* for stock i at time interval t. Consistent with other studies (such as Antweiler & Frank; Kim & Kim, 2014), 1 is added in the log-transformed equation as indicated above to cater for firms with no *tweets* at any point in time since the natural logarithm of zero is undefined. Log-transformed aggregate *tweets* are used as this allows for computing elasticities (since some dependent variables are log-transformed) and controls for scaling.

The message agreement index ($A_{i,t}$) reflects the extent of the consensus among microbloggers on the prospects of each stock i at time-variable t. Following Antweiler and Frank (2004), the following is used to measure message agreement:

$$A_{i,t} = 1 - \sqrt{1 - \left(\frac{Pos_{i,t} - Neg_{i,t}}{Pos_{i,t} - Neg_{i,t}}\right)^2} \in [0,1] \qquad \text{Equation [3]}$$

Where $Pos_{i,t}\ and\ Neg_{i,t}$ indicate the number of messages which are respectively categorised as positive and negative. If all the messages at a given time are equally distributed between positive and negative, it means that there is absolute agreement among microbloggers and therefore the value of A_t will be equivalent to 1. In a situation where all messages are either positive or negative, then it means microbloggers are in total disagreement and the agreement index will therefore be 0. Thus, the closer the agreement index is to 0, the greater the disagreement among stock microbloggers while a value close to 1 indicates greater agreement. If disagreement encourages trading as hypothesised, then A_t should be inversely associated with measures of the trading volume. The major advantage of the above agreement metric is that it directly measures the dispersion of investor opinions compared to alternative metrics which rely on indirect measures like volatility and analyst forecast dispersion. Additionally, the agreement measure is computed at the daily level compared to alternative metrics which are usually measured at lower monthly and

quarterly frequencies (Diether, Malloy & Scherbina, 2002). The challenge with the agreement index above is that there are companies that go for a considerable time without being mentioned on Twitter and StockTwits forums, leading to missing values for the *tweet* features. Consistent with Antweiler and Frank (2004) and Sprenger *et al.*, (2014), this study assigns a value of zero to *tweets* features for all "quiet" periods. Imputing zero values to companies that are not mentioned on the Twitter and StockTwits platforms on any day is done on the presumption that when investors do not mention a specific counter, this means that they are neutral on the prospects of the counter.

3.2.2 Market Features

Three stock market features are used in the study, namely; raw returns ($R_{i,t}$), Volatility $V_{i,t}$ and Trading Volume $TV_{i,t}$. Raw returns ($R_{i,t}$) for stock i at time interval t are defined as:

$$R_{i,t} = ln\left(\frac{P_t}{P_{t-1}}\right) \qquad \text{Equation [4]}$$

Where P_t stands for the closing price at time interval t and P_{t-1} stands for the closing price at time interval $t-1$. Continuously compounded returns are used instead of discrete returns to overcome any issues of non-normality since stock returns are usually assumed to be lognormally distributed (Brooks, 2019). The continuously compounded formula for computing returns in Equation [4] is therefore taken to constitute a non-linear transformation of the data. For holidays when there is no trading on the JSE, missing values are imputed using the average of the closing price a day before and a day after the missing value day in conformity with previous studies (such as Sprenger *et al*., 2014). The missing values of the closing prices are therefore imputed using the following formula:

$$m_{i,t} = \frac{P_{i,t-1} + P_{i,t+1}}{2} \qquad \text{Equation [5]}$$

Where $m_{i,t}$ represents the imputed value of the closing price of stock i at time t, $P_{i,t-1}$ is the closing price of stock i the day before the missing value and $P_{i,t+1}$ is the closing price of stock i the day after the missing value. Since holidays represent a small percentage (less than 5%) of the observations, this is not expected to have a significant effect on the findings.

In their study, Antweiler and Frank (2004) used the generalised autoregressive conditional heteroskedasticity (GARCH) class of volatility models as proxies for volatility. Though the GARCH family of models cater for volatility clustering and are useful for approximating and modelling temporal conditional volatility, they are considered inefficient and inaccurate because they rely on closing prices and are prone to error especially during turbulent days (Chou, Chou & Liu, 2015). To ameliorate the weaknesses of the GARCH family of models outlined above, this study, therefore, uses a volatility estimation model that captures drops and recoveries of financial markets daily instead of the classical close-to-close volatility models. Also, GARCH models are useful in capturing the implied volatility and the current study is not interested in measuring implied volatility, and hence GARCH models are not applicable. Volatility ($V_{i,t}$) in this study is estimated using intraday price highs and lows in line with the PARK volatility measure (Parkinson, 1980). The PARK estimator's accuracy instinctively emanates from the idea that the intraday price range gives more information regarding future volatility than two arbitrary closing-price points in the series. Supposing that the stock price follows a simple Brownian model without a constant term, the PARK statistic is calculated as follows:

$$VOL^{PARK} = \frac{(\ln(H_t - \ln(L_t)))^2}{4\ln(2)} \qquad \text{Equation [6]}$$

Where $H_t \; and \; L_t$ stand for the daily highs and lows of a stock price. PARK volatility has also been used in scholarly articles examining the role of textual sentiment in capital markets (such as Sprenger *et al*., 2014; Li, van Dalen & van Rees, 2018) and is, therefore, useful for comparing the findings from this study with previous studies. Trading volume ($TV_{i,t}$) is estimated as the natural logarithm of the traded volume of stock i at time t. The log-transformed trading volume is

calculated using $ln(1 + trading\ volume)$ to cater for firms with a zero trading volume at any time since the natural logarithm of zero is undefined.

3.3 Econometric model estimation

3.3.1 Contemporaneous relationship between *tweet* features and stock market features

This subsection presents the tests for the primary hypotheses of the study. Following previous studies (such as Twedt & Rees, 2011; Sprenger *et al.*, 2014), panel regressions with ticker fixed-effects are used to examine the contemporaneous link between *tweet* features and market features as shown in Equation [7]. This model follows that of Sprenger *et al.*, (2014); Agrawal, Azar, Lo and Singh (2018) where all the *tweet* features are used as covariates and the market index is included as a control variable as shown below:

$$Y_{i,t} = \beta_1 B_{i,t} + \beta_2 M_{i,t} + \beta_3 A_{i,t} + \beta_4 R^m{}_t + \delta_i + \epsilon_{i,t} \qquad \text{Equation [7]}$$

Where

$Y_{i,t}$ represents the three market features (firm-level stock returns, volatility and trading volume) of firm i at time t;

$B_{i,t}$ is bullishness

$M_{i,t}$ represents message volume of firm i at time interval t;

$A_{i,t}$ represents message agreement for firm i at time interval t;

$R^m{}_t$ is the market return calculated as the natural logarithm of the closing value of the JALSH index at time t divided by the value at time $t-1$;

$\epsilon_{i,t}$ is an error term that is clustered by ticker;

δ_i is the unknown intercept for every company

3.3.2 Pooled quantile regressions

To examine the link between the magnitude of *tweet* features and stock market features, the study uses quantile regression[5]. Quantile regressions are used in this study to test the secondary hypotheses of the study. This type of model was first proposed by Koenker and Bassett (1978) and

[5] This is estimated using the Rstudio Package *"quantreg"*

has been used widely in studies examining investor sentiment as well as in asset pricing studies (such as Swamy & Dharani, 2019). While some studies have used the sample splitting procedure to examine the magnitude and significance of investor sentiment at different market conditions (such as Li, Shen & Zhang, 2019), Koenker (2004) argues that the procedure leads to severe sample selection complications. Quantile regression necessitates the approximation of conditional quantiles of the regressand given a range of predictor variables without splitting the sample (Koenker & Bassett, 1978). The quantile regression estimator also permits the effect of the explanatory variable to fluctuate across quantiles of the dependent variable.

Some of the documented advantages of quantile regression estimators include the fact that they are robust to outliers (Koenker & Machado, 1999) and they deal with non-linearity without presuming a specific form of the model (Koenker & Bassett, 1978). Quantile regression is also useful because it allows the comparison of coefficients across quantiles using interquartile regression (Koenker & Bassett, 1978). Using quantile regression, the conditional quantile function of $\gamma_{i,t}$ at quantile τ given explanatory variable $x_{i,t}$ is defined as follows:

$$Q_\tau(\gamma_{i,t}|x_{i,t}) = c_\tau + \beta_\tau x_t + F_{\varepsilon_t}^{-1}(\tau) \qquad \text{Equation [8]}$$

Where F_ε stands for the distribution of errors and β_τ and c_τ are the parameters. The coefficients of the τ^{th} conditional quantile regression are approximated as follows:

$$\hat{\beta}^\tau = \underset{c_\tau,\ \beta_\tau \in \mathbb{R}}{\arg\min} \sum_{t=1}^{T-1} \rho_\tau \left(\gamma_{i,t} - \left(c_\tau + \beta_\tau x_{i,t}\right)\right) \qquad \text{Equation [9]}$$

Where T indicates the sample size and ρ_τ is the check function defined as $\rho_\tau(\varepsilon) = \tau\varepsilon$ if $\varepsilon \geq 0$ and $\rho_\tau(\varepsilon) = (\tau - 1)\varepsilon$ otherwise. This study uses Equation [7] as the baseline equation for the quantile regression model specification as follows:

$$Y_{it,\tau} = \alpha_{\tau} + \beta_1 B_{it,\tau} + \beta_2 M_{it,\tau} + \beta_3 A_{it,\tau} + \beta_4 R^m{}_{t,\tau} + \epsilon_{it} \qquad \text{Equation [10]}$$

where τ is the τ^{th} quantile in the conditional distribution of the regressand. Equation [10] is the quantile regression model which is estimated to establish whether the magnitude of sentiment is linked to the magnitude of stock return. Following previous studies (such as Swamy & Dharani, 2019), five quantile intervals representing the different market conditions as shown by the firm-level stock returns are used as follows: $\tau \in (0.1, 0.25, 0.5, 0.75, 0.9)$ where $\tau \in (0.1, 0.25)$ represent bad market conditions, $\tau \in (0.5)$ represents normal market conditions and $\tau \in (0.75, 0.9)$ represent good market conditions. For robustness, the model is re-estimated using $\tau \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$ and $\tau \in (0.05, 0.2, 0.4, 0.6, 0.8, 0.95)$.

3.3.3 Granger non-causality tests

This subsection presents the tests used to test the tertiary hypotheses of the study. To examine whether past values of *tweet* features contain useful information that could be used to predict future values of stock returns, the study utilises Granger non-causality tests (Granger, 1969). According to Granger (1969), a causal relationship is inferred when the lagged values of a variable X_t have explanatory power in a regression of a variable Y_t on lagged values of X_t *and* Y_t. Though Granger causality does not mean causality in the philosophical sense of the word, it is used to infer if past values of a variable contain useful information that can be used to predict future values of another variable. To test whether *X* (representing the *tweets* features) Granger-causes *Y* (representing stock returns), the Wald statistic for heterogenous panels proposed by Dumitrescu and Hurlin (2012) is used. The test statistic Dumitrescu and Hurlin (2012) propose is a simple average of individual Wald statistics obtained from testing the null hypothesis for every cross-sectional unit in the panel. The test is based on the stationary fixed-effects panel model:

$$Y_{i,t} = \alpha_i + \sum_{k=1}^{K} \beta_{i,k} Y_{i,t-k} + \sum_{k=1}^{K} \gamma_{i,k} X_{i,t-k} + \varepsilon_{i,t} \qquad \text{Equation [11]}$$

Where $X_{i,t}$ and $Y_{i,t}$ are the *tweets* features and the stock returns respectively. The maximum number of lags is restricted to two in line with previous studies. For robustness, the optimal lag length is

also selected using the lag length that minimises the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan-Quinn Information Criterion (HQIC). The *"homogenous non-causality"* null hypothesis of the Dumitrescu and Hurlin (2012) statistic is given below:

$$H_0: \gamma_{i1} \dots \gamma_{iK} = 0 \quad \forall i = 1, \dots N$$

Where it is assumed that there is no causal relationship under the null hypothesis for all N while $N - N_1$ causal relationships are assumed under the alternative hypothesis where $N_1 < N$. N_1 is assumed to be unknown and will comply with the condition $0 \leq N_1/N < 1$. Dumitrescu and Hurlin (2012) demonstrate that under the null hypothesis, the panel Wald statistic ($Z_{N,T}^{Hnc}$) is asympotically distributed according to a normal distribution with a mean of 0 and variance of 1 as T approaches infinity as follows:

$$Z_{N,T}^{Hnc} = \sqrt{\frac{N}{2K}}(W_{N,T}^{Hnc} - K) \qquad \text{Equation [12]}$$

Where $W_{N,T}^{Hnc} = (1/N)\sum_{i=1}^{N} W_{i,T}$. $W_{i,T}$ is the individual Wald test statistic for i^{th}cross-sectional unit corresponding to the individual test $H_0: \gamma_i = 0$ and is calculated as follows:

$$W_{i,t} = \widehat{\theta'}_i R'[\hat{\sigma}_i^2 R(Z'_i Z'_i)^{-1} R']^{-1} R\hat{\theta}_i = \frac{\hat{\theta}_i' R'[R(Z'_i Z'_i)^{-1} R']^{-1} R\hat{\theta}_i}{\hat{\varepsilon}_i' \hat{\varepsilon}_i'/(T - 2K - 1)} \qquad \text{Equation [13]}$$

Where $\hat{\theta}_i$ is the estimate of the parameter θ_i obtained under the null hypothesis and σ_i^2 is the estimate of the variance of the residuals and $Z_i = [e: Y_i: X_{1i}]$ is a $(T, 2K + 1)$ matrix constructed by a horizontally concatenated $(T, 1)$ unit vector, a (T, K) matrix Y_i and a (T, K) matrix X_{1i}.

3.4 Robustness checks

To mitigate methodological choices from driving the results, some econometric models above are re-estimated using a multitude of alternative specifications which are outlined in this section.

3.4.1 Random effects regression

To test the robustness of the fixed effects model used to examine the relationship between *tweet* features and stock market features (Outlined in Section 3.3.1), a panel regression model with firm-specific random effects (μ_i) is also estimated. Compared to the fixed-effects model, a panel regression estimation with firm-specific random effects allows parameter estimates to be affected by variations in both cross-sectional and time-series aspects of the data (Antweiler & Frank, 2004). Following Antweiler and Frank, (2004), the fixed-effects model in Equation [5] is reestimated using company-specific random effects as follows:

$$Y_{i,t} = \beta_1 B_{i,t} + \beta_2 M_{it} + \beta_3 A_{i,t} + \beta_4 R^m{}_t + \mu_i + \mu_{i,t} \qquad \text{Equation [14]}$$

Where the variables are as previously defined in Section 3.3.1

3.4.2 Quantile regressions with fixed effects

The pooled quantile regression approach outlined in Section 3.3.2 might be susceptible to the problem of unobserved heterogeneity. Koenker (2004) proposes a fixed-effects quantile regression model that addresses the possibility of endogenous predictor variables. Consistent with Koenker (2004), the following model for the conditional quantile functions of the dependent variable of firm i at time t is used:

$$Q_{yi,t}(\tau|x_{i,t}) = \alpha_i + x'_{i,t}\beta(\tau) \qquad \text{Equation [15]}$$

Where α_i indicates the firm fixed effects, $x'_{i,t}$ is a vector of predictor variables and the $\tau - dependent$ vector β is the vector of predictor variables of parameters to be estimated. In the model, the firm fixed effects α_i imply a pure location shift on the conditional quantiles of the regressand. Effectively, the effects of the predictor variables are allowed to be dependent on the quantile τ while the effects of α_i are not, but are still useful to control for unobserved heterogeneity. To estimate the model in Equation [15], Koenker (2004) proposes solving the following using linear programming:

$$\min_{\alpha,\beta} \sum_{k=1}^{q} \sum_{t=1}^{T} \sum_{i=1}^{n} \omega_k \, \rho_{T_k}(y_{i,t} - \alpha_{i,t} - x'_{i,t}\beta(\tau_k)) \qquad \text{Equation [16]}$$

Where

ω_k denotes the weights controlling the relative influence of the quantiles on the estimation of the α_i parameters,

$\rho_\tau(\,.\,)$ is the quantile loss function.

The fixed effects panel quantile regression methodology outlined herein has been widely used in various studies in the area of finance, particularly studies looking at investor sentiment and other behavioural finance topics (such as Swamy & Dharani, 2019). This study estimates the fixed effects panel quantile regression model in Equation [15] using the *"rqpd"* Rstudio package to optimally solve Equation [16].

3.4.3 Lag-lead relationships

To test the robustness of the Granger non-causality tests outlined in Section 3.3.3, the study also uses the Fama and Macbeth (1973) style regressions to test the predictive power of *tweet* features on stock returns. For robustness, previous studies have used time sequencing regressions as well as Fama Macbeth style regressions to find the temporal structure of the relationship between *tweet* features and stock market features. Though some studies (such as Antweiler and Frank, 2004) analysed the temporal structure of the *tweet* and market features using time-sequencing regressions, Sprenger *et al.,* (2014) argue that this model has disadvantages. These advantages include, among other things, the potential existence of a time trend that could explain the results and the need to use clustered errors to account for the lack of independence in the stock returns. To circumvent the above-mentioned disadvantages of time sequencing regressions, this study, therefore, follows Sprenger *et al.*, (2014) and Kim and Kim (2014) by employing the Fama and MacBeth (1973) cross-sectional regressions.

In estimating the Fama-Macbeth regression, the two pass-regression process firstly estimates the cross-sectional regressions of the firm-level market features and stock returns for each day. This is followed by estimating the averages of the daily coefficients and reporting the $t\ statistics$ of

the estimates. Heteroskedasticity-consistent standard errors are reported to deal with heteroskedasticity and autocorrelation concerns of dealing with panel data. Like previous studies, the study controls for size using the natural logarithm of the firms' market capitalisation. One and two-day lags of *tweet* features are regressed on stock returns separately and *vice-versa* to establish the direction of prediction. The Fama and MacBeth (1973) style regressions used for investigating the temporal structure of the variables is shown in Equation [17] below:

$$Y_{i,t+j} = \sigma_t + \beta_t B_{i,t} + \gamma_t M_{it} + \delta_t A_{i,t} + \beta_t Size_{i,t} + \epsilon_{i,t} \qquad \text{Equation [17]}$$

Where:

$Y_{i,t+j}$ represents the market features for the future ahead cases of $j = 1, 2$;

$B_{i,t}$, M_{it} and $A_{i,t}$ represent bullishness, message volume and agreement; and

$Size_t$ is a control variable defined as the natural logarithm of the market value of stock i at time t.

3.5 Summary

This section has outlined the data and methodological specifications for testing the research hypotheses identified in the introductory section of this study. The methods and variables were put into the context of other related studies done in developed and emerging markets. Quantile regression, which has been rarely used in behavioural finance studies in a South African context is introduced to test whether there is a monotonic relationship between the magnitude of *tweet* features and stock market features. The methods specified for this study are generally in line with previous studies done. Alternative methodological considerations are also explored to make sure that the results from the study are not driven by the choice of the method used. The next section presents the results from the methodological specifications outlined in this section.

4 RESULTS AND DISCUSSION

This section outlines the findings from the methodology explained in the preceding section. Firstly, the section presents the descriptive statistics for the variables used in the study, followed by the results from the specific hypotheses developed for the study.

4.1 Descriptive statistics

Table 1 below shows the summary statistics of the market and microblog features for the daily granular data used in the study. The summary statistics are for the 158 companies[6] that formed part of this study. The panel data includes observations of the *tweet* and market features for a period of 1247 trading days from 2015 to 2019, together making a total of 197 026 firm-days.

Table 1: Summary statistics of the variables

Statistic	N	Mean	St.Dev	Med	Min	Max
Tweet features						
Bullishness	197 026	0.0020	0.1035	0.0000	-0.0069	0.9978
Messages	197 026	0.4133	0.9295	0.0000	0.0000	7.8860
Agreement	197 026	0.0212	0.0605	0.0000	0.0000	0.8549
Market features						
Stock returns	197 026	-0.0001	0.0232	0.0000	-0.9525	0.4964
Trading volume	197 026	13.0393	1.8832	13.34898	0.0000	28.7688
Volatility	197 026	26.4899	8.1698	26.2757	6.8702	56.5512
Size	197 026	23.8417	1.4584	23.6327	17.4128	28.7688
Market return	197 026	0.0001	0.0093	0.00046	-0.03621	0.03649

The variables displayed in Table 1 are as previously defined in Section 3.2 while the Trading volume, message volume and size are taken as the natural logarithms of the trading volume, message volume and the market capitalisation of the sampled firms respectively. As shown in Table 1, bullishness ranges from a minimum value of -0.9969 to a maximum value of 0.9978 showing that textual sentiment as measured by bullishness fluctuates across the polar estimates. On average, bullishness is +0.002, which is a positive figure but close to 0 showing that in the

[6] A complete list of all the firms which were included in the study is included in Appendix A

study period, the sentiment of investors was only slightly positive and therefore near neutral. The distribution of the number and percentage of firm-level *tweets* by day of the week is visually depicted in Panels A and B of Figure 2 respectively.

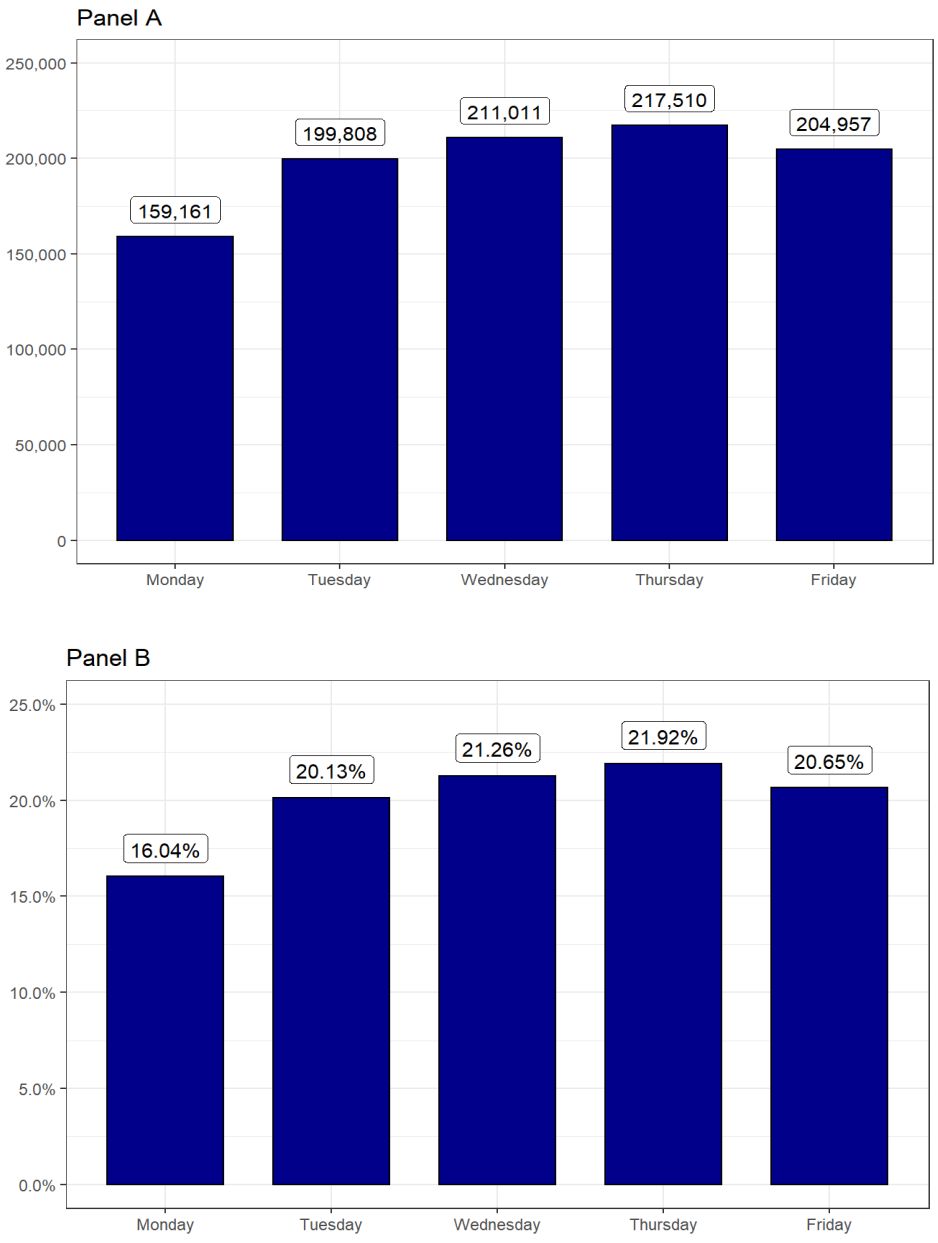


Figure 2: Distribution of *tweets* by day of the week

As shown in Figure 2, the total number of times the sampled companies were mentioned on the Twitter and StockTwits platforms is 992 447. The distribution of the *tweets* by day of the week shows that the volume of stock-related messages is low at the beginning of the week and gradually increases until it peaks on Thursdays and thereafter subsides. Figure 3 visualises the daily distribution of the aggregate *tweets* for the sample period.

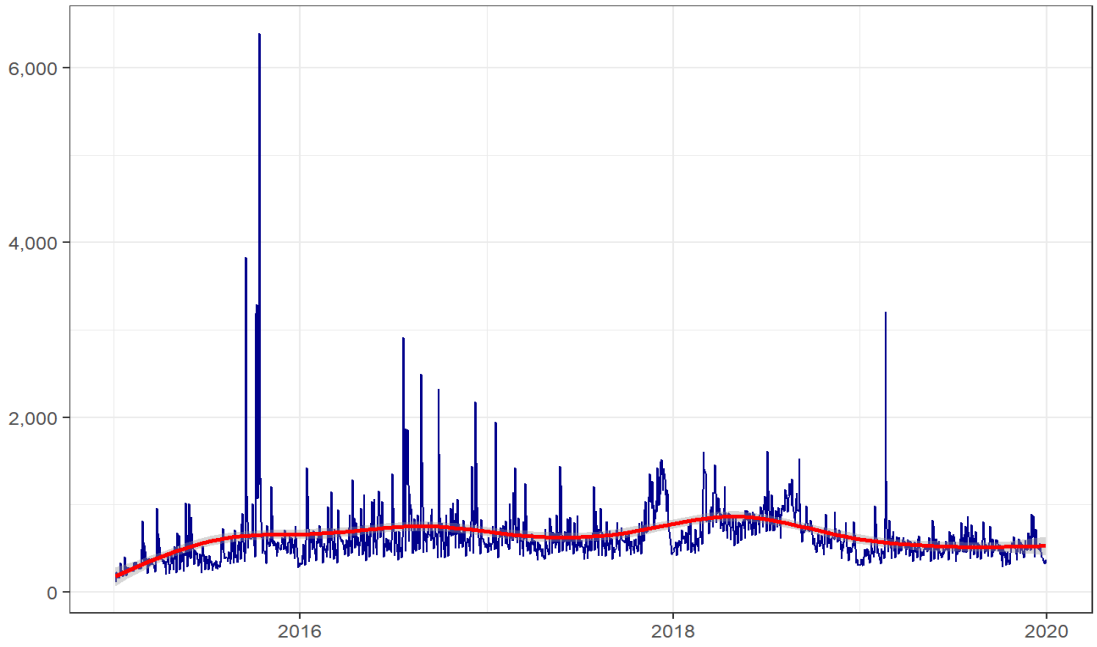
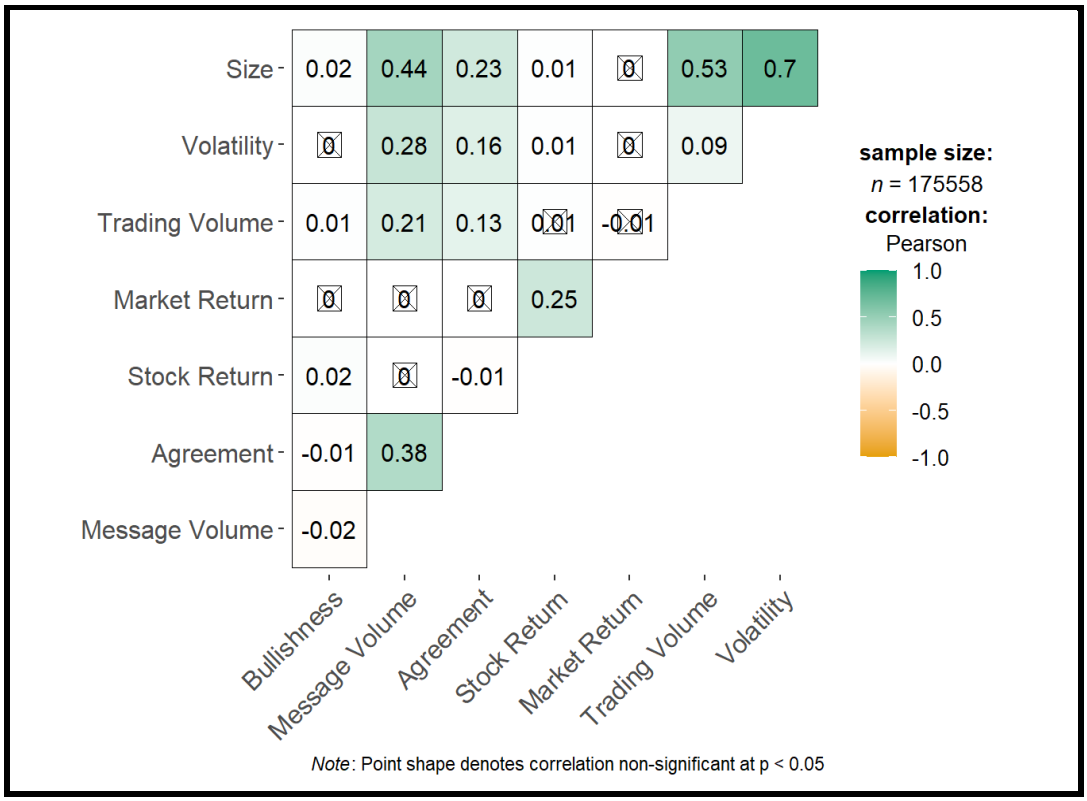


Figure 3: Daily distribution of the aggregate *tweets*
Note: *The blue line shows the time series of the daily aggregated tweets, the red line shows the smoothed conditional means line with 95% confidence bands*

As shown in Figure 3, on average, the aggregate number of *tweets* posted daily generally ranges between 0 and 2000 *tweets* with intermittent spikes in between; for example, the spike in *tweets* amounting to 6 399 *tweets* on a single day in 2015. To mitigate the influence of outliers, following previous studies, the *tweet* features are winsorised at the 5% level (2.5% from both ends). The initial links between *tweet* features and the stock market features are depicted by the simple pairwise correlation matrix in Table 2:

Table 2: Correlations among tweet features and market features



The pairwise Pearson Correlations between the alternative *tweet* features (bullishness, message volume and agreement), are generally low showing that there are some somewhat relationships among the *tweet* features, while also demonstrating that the metrics capture different aspects of textual sentiment. The correlation coefficients between *tweet* features and market features show some significant associations for some variables (such as between message volume and trading volume; agreement and stock returns) while some correlations between *tweet* features and stock returns (message volume and stock returns; volatility and bullishness) are insignificant. The low to moderate correlation values (all less than 0.7) in the correlation matrix also demonstrate the absence of potential multicollinearity in the variables. The significance of most of the correlation values postulates that they may produce signals worth investigating more formally by utilising rigorous econometric models to unveil further links that may be present. The results from the rigorous econometric models are shown in the following section in line with the study hypotheses.

4.2 Contemporaneous relationship between *tweet* features and market features

This section presents the results on the contemporaneous associations between the *tweet* features and market features to establish whether *tweets* can explain the cross-section of market features for JALSH constituent companies. To determine the model to use between the fixed effects model and the random-effects model, the Hausman test was used, where the null hypothesis is that the latter model is the preferred model. The Hausman test determines whether the unique errors μ_i are correlated with the regressors. Both the fixed effects and the random-effects model were estimated and results subjected to the Hausman test. The results from the Hausman test are displayed in Table 3 below:

Table 3: Results from the Hausman test

	Dependent variable		
	Return	*Trading volume*	*Volatility*
χ^2	332.48	44.482	8.6701
$p-value$	0.0000	0.0000	0.0499

Null Hypothesis: Preferred model is random effects

The results in Table 3 show that for the three models used to examine the contemporaneous relationship between *tweet* features and market features, the coefficients are all statistically significant at the 5% level of significance. This means that for all the models, the null hypothesis of preference for the random-effects model can be rejected in favour of the fixed-effects model. The fixed effects[7] models are therefore used to formally examine the contemporaneous relationships between the *tweet* features and the market features in this study. This is in line with the majority of previous studies that used panel data (such as Antweiler & Frank, 2004; Twedt & Rees, 2011). The results from the contemporaneous OLS estimations with firm-fixed effects are shown in Table 4:

[7] For robustness, the results for the random effects models are included in Table C1 of Appendix C. The results are qualitatively similar to the results obtained in this section using the fixed effects models

Table 4: Contemporaneous OLS regressions with firm fixed effects

	Dependent Variable		
	Return	*Volatility*	*Trading volume*
	(1)	*(2)*	*(2)*
Bullishness	0.004***	1.086***	-0.000
	(0.001)	(0.046)	(0.026)
Agreement	-0.004***	-1.647***	0.209***
	(0.001)	(0.084)	(0.047)
Messages	0.0001	-0.208***	0.083***
	(0.0001)	(0.008)	(0.004)
Market Return	0.615***	0.450	-0.867***
	(0.006)	(0.467)	(0.262)
Constant	-0.001	0.001	-0.002
	(0.003)	(0.003)	(0.003)
Observations	180 724	180 887	180 881
R^2	0.062	0.011	0.002
Adjusted R^2	0.061	0.010	0.002
F Statistic	2.997.6***(df=4)	500.5***(df=4)	108.4***(df=4)

Notes:
1. *The first row for each combination of tweet feature and market feature reports the regression coefficients, the second row reports the standard errors (in brackets)*
2. **p<0.1; **p<0.05; ***p<0.01*

In Table 4, Models 1, 2 and 3 show the regressions of returns, volatility and trading volume on the *tweet* features respectively. The findings shown in the table reveal that bullishness is directly and significantly associated with stock returns. In the results, quality and content seem to be more vital than quantity since bullishness is related to returns more strongly than message volume. The relationship between the agreement index and stock returns is negative and statistically significant at 1% ($\beta = -0.004, p < 0.01$) showing that disagreement among microbloggers is associated with higher stock returns. This is in line with previous studies which report a negative and significant relationship between investor agreement using texts extracted from Yahoo! Finance (Antweiler & Frank, 2004) and Twitter (Sprenger *et al*., 2014; Li, van Dalen & van Rees, 2018). The relationship between message volume and stock returns, though positive ($\beta = 0.0001$), is not significant at any of the conventional levels. This is consistent with Sprenger *et al*. (2014) who found no significant relationship between the natural logarithm of the volume of firm-level *tweets* and firm-level returns. These results are however contrary to the findings of Antweiler and Frank (2004) and Li, van Dalen and van Rees (2018) who found a positive and significant relationship between message volume and stock returns.

The results presented in Table 4 also show that volatility is significantly associated with all the *tweet* features. Firstly volatility is positively and significantly associated with bullishness ($\beta = 1.086, p < 0.01$). This is consistent with previous studies which report a positive and significant relationship between volatility and bullishness (Kim & Kim, 2014; Sprenger *et al*., 2014). As expected, volatility is negatively and significantly associated with the agreement ($\beta = -1.647, p < 0.01$). This means that increased dispersion of investor opinion on stock microblogs is linked with higher volatility. Volatility is also negatively and significantly associated with the volume of firm-level *tweets* ($\beta = -2.08, p < 0.01$). This is contrary to the majority of previous studies which largely report a significant and direct link between message volume and volatility (Antweiler & Frank, 2004; Twedt & Rees, 2011; Sprenger *et al*., 2014).

On the contemporaneous relationship between trading volume and the *tweet* features, in line with Li, van Dalen and van Rees (2018), Table 4 shows that there is no statistically significant relationship between bullishness and trading volume. On the other hand, the link between message volume and trading volume is positive and statistically significant at the 1% level ($\beta = 0.083, p < 0.01$). Since the values for message volume and trading volume are both log transformed, they can be interpreted as elasticities. This means that a 1% increase in message volume is associated with a more than 8% increase in trading voume. This result is almost quantitively similar to Sprenger *et al*., (2014) who report that a 1% increase in message volume is associated with a 10% increase in trading volume. The positive relationship between message volume and trading volume means that microblog users on Twitter post messages of companies that are traded more heavily (Sprenger *et al*., 2014). Contrary to previous studies (such as Antweiler & Frank, 2004; Li, van Dalen & van Rees, 2018), the relationship between the agreement index and trading volume shown in Table 4 is positive. This is in line with literature that recognises that information differences need to interact with some other forms of heterogeneity, like heterogeneous beliefs, to generate trading (Sprenger et al., 2014). However, the extent to which how much each source of disagreement matters for trading is an open question.

4.3 The relationship between the magnitude of stock returns and *tweet* features

To test the relationship between the magnitude of stock returns and *tweet* features, pooled quantile regressions[8] were used in line with the secondary hypotheses. The empirical results from the quantile regressions at the specified return quantiles ($\tau \epsilon [0.1, 0.25, 0.5, 0.75, 0.9]$) are presented in Table 5.

Table 5: Pooled quantile regression estimates[9]

	Dependent variable:				
	Stock Return				
	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
Bullishness	0.005***	0.003***	0.002***	0.001*	0.001
	(0.001)	(0.001)	(0.0004)	(0.001)	0.001)
Agreement	-0.020***	-0.008***	-0.001*	0.005***	0.018***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Messages	-0.001***	-0.000***	0.0001***	0.0004***	0.001***
	(0.000)	(0.0001)	(0.0001)	(0.000)	(0.000)
Market return	0.651***	0.603***	0.501	0.603***	0.664***
	(0.010)	(0.006)	(0.004)	(0.006)	(0.010)
Constant	-0.022***	-0.010***	-0.0004***	0.009***	0.022***
	(0.000)	(0.000)	(0.00004)	(0.000)	(0.000
Pseudo R^2	0.042	0.045	0.031	0.043	0.038
Observations	180 724	180 724	180 724	180 724	180 724

Notes:
1. *The first row for each return quantile reports the regression coefficients for each variable, the second row reports the standard errors (in brackets)*
2. **p<0.1; **p<0.05; ***p<0.01*

Firstly, the relationship between stock returns and the control variable, market returns, is statistically significant at the lower and upper return quantiles while at the middle quantiles the relationship is insignificant. It can also be noted that the magnitude of the coefficient starts high at the low return quantiles, falls until reaching a minimum at the middle return quantiles before rising again in the upper return quantiles. The magnitude and significance of the control variable show that the association between firm-level stock returns and market returns is stronger and more significant during good times and bad times while the relationship is weak and insignificant during

[8] For robustness, the quantile regressions are re-estimated with firm-fixed effects in line with the methodology specified in Section 3.4.2. The results obtained from the model specification estimated using the *"rqpd"* R package are shown in Table C2 in Appendix C. The results are consistent with the results obtained using the pooled quantile regression. The results are therefore robust to the inclusion of firm fixed effects.

[9] The R-code used to get the estimates is attached as Appendix D in the study.

normal times. The results showing the magnitude of the relationship between *tweet* features and stock returns are discussed below:

4.3.1 Bullishness and stock returns

Table 5 shows that bullishness is positively associated with stock returns at all quantiles of stock returns. However, the magnitude of the relationship between bullishness and stock returns persistently falls as we move from the lowest quantile of returns ($\tau \in [0.1]$) to the highest quantile of returns ($\tau \in [0.9]$). This is seen as the coefficient of the relationship falls gradually from $\beta = 0.005$ in the lowest return quantile to $\beta = 0.001$ in the highest return quantile. Moreover, the relationship between bullishness and returns becomes less statistically pronounced as we move from the lowest quantile of returns to the highest. The relationship between bullishness and stock returns at the low and middle return quantiles ($\tau\epsilon[0.1, 0.25, 0.5]$) is statistically significant at the 99% confidence level and the significance falls to 90% confidence level at $\tau\epsilon[0.75]$ while the relationship at the highest quantile $\tau\epsilon[0.9]$ is not statistically significant at any of the conventional confidence levels. The relationship between bullishness and stock returns at all the return quantiles is visualised in Figure 4:

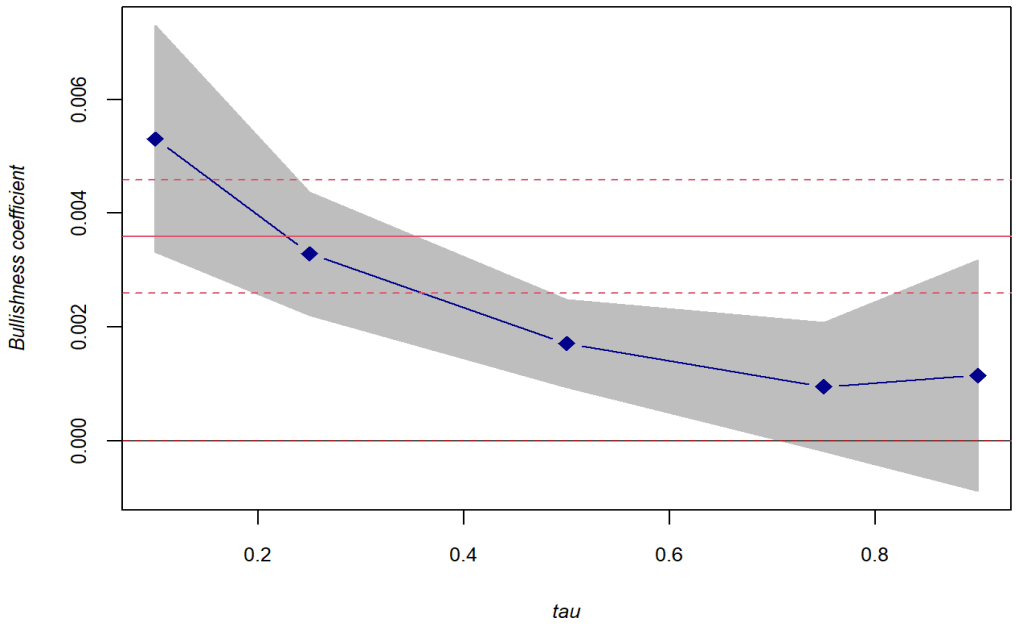


Figure 4: Bullishness - QR and OLS estimates with the 95% confidence intervals

Note: *The red continuous line shows the Ordinary Least Squares (OLS) estimate while the two dotted lines show the 95% confidence interval of the OLS estimates and the brown continuous line shows the point where the vertical axis is equal to zero. The vertical axis shows the coefficients for the bullishness score at the various quantiles while the horizontal axis indicates the quantile distributions of the returns* ($\tau\epsilon[0.1, 0.25, 0.5, 0.75, 0.9]$. *The values of the estimated* $\beta(\tau)$ *parameters are connected by the blue line while the grey shaded area indicates the 95% confidence intervals of the* $\beta(\tau)$ *estimated parameters.*

Figure 4 also shows that bullishness falls by increasingly higher margins as we move from lower quantiles of the return distribution but flatten out at higher return quantiles. The results presented in Table 5 and visualised in Figure 4 also show that textual sentiment extracted from *tweets* (bullishness) can explain stock market returns under bad and normal market conditions ($\tau\epsilon[0.1, 0.25, 0.5]$) but not under good conditions ($\tau\epsilon[0.75, 0.9]$). This is in line with Allen, McAleer and Singh (2019) who report significantly stronger relationships between textual sentiment extracted from online news articles and stock returns at lower quantiles of the latter. This can be accounted for by more investor decisions being based on stock microblogs around a time when the stock prices are declining than when stock prices are on the rise.

Table 5 also reveals that the use of traditional optimisation techniques like the Ordinary Least Squares to examine the relationship between stock returns and textual sentiment generally provides

biased estimates. The OLS estimator only coincides with quantile estimators at $\tau \in [0.25]$ and for the remainder of the return quantiles, the coefficients of the bullishness index are outside the 95% confidence interval bands of the OLS estimates. Generally, the results show a monotonic relationship between the magnitude of bullishness and stock returns up to $\tau \in [0.5]$ with the relationship becoming flat at the two highest return quantiles $\tau \in [0.75, 0.9]$. The null hypothesis of a monotonic relationship between the magnitude of bullishness and the magnitude of stock returns can therefore be rejected.

4.3.2 Investor agreement and stock returns

The results on the relationship between investor agreement and stock returns shown in Table 5 show three patterns emanating from the quantile distribution of returns. Firstly, the $\beta(\tau)$ estimates increase in absolute terms as we move outwards from the middle quantile ($\tau\epsilon[0.5]$). to the extreme return quantiles ($\tau\epsilon[0.1,\ 0.25)$ and ($\tau\epsilon[0.75, 0.9]$). Secondly, the coefficients are more significant ($p < 0.01$) at lower return quantiles ($\tau\epsilon[0.1, 0.25]$) and higher return quantiles ($\tau\epsilon[0.75, 0.9]$) while the relationship is less significant ($p < 0.1$) at the middle of the return distribution ($\tau\epsilon[0.5]$). Thirdly, the relationship between investor agreement and returns is negative (positive) at low $\tau \in [0.1, 0.25, 0.5]$ (high) $\tau\epsilon[0.75, 0.9]$ return quantiles. This relationship is visualised in Figure 5:

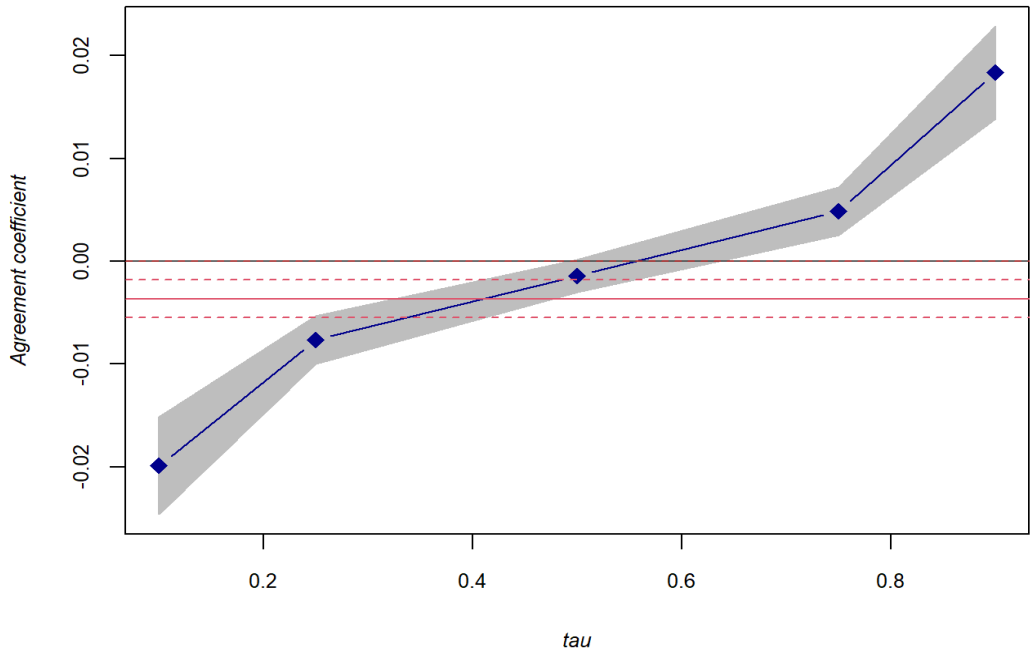


Figure 5: Agreement - QR and OLS estimates with the 95% confidence intervals

Note: *The red continuous line shows the Ordinary Least Squares (OLS) estimate while the two dotted lines show the 95% confidence interval of the OLS estimates and the brown continuous line shows the point where the vertical axis is equal to zero. The vertical axis shows the coefficients for the agreement index at the various quantiles while the horizontal axis indicates the quantile distributions of the returns* $(\tau\epsilon[0.1, 0.25, 0.5, 0.75, 0.9]$. *The values of the estimated* $\beta(\tau)$ *parameters are connected by the blue line while the grey shaded area indicates the 95% confidence intervals of the* $\beta(\tau)$ *estimated parameters.*

Figure 5 shows that the coefficients (in absolute terms) of the dispersion of investor opinions as measured from *tweet* messages increases as we move outwards from middle return quantiles to extreme return quantiles. As the quantile levels move up, the estimates of the agreement index vary widely in sign, magnitude and significance. The results are in line with the three hypotheses suggested by Diether, Malloy and Scherbina (2002) on the association between investor disagreement and stock returns. The authors ascribe the association between agreement and return to three statuses of overpricing correction process and hypothesise (i) an inverse relationship between agreement and returns when the overpricing is corrected (ii) a positive association between the variables when the overpricing is continuing and (iii) a trivial relationship when the overpricing process is completed. Figure 5 demonstrates three statuses of mispricing correction as the relationship between the agreement index and returns is negative at lower quantiles, positive at upper quantiles and trivial at the middle quantile.

According to Cujean and Hasler (2017), in bad times, news exacerbates disagreement by polarizing agents' opinion while in good times, the reaction of disagreement to news is weak and exhibits little persistence. This spike in disagreement in bad times creates positive serial correlations in returns at short horizons. This is shown in Figure 5 which shows a negative relationship between investor agreement and returns at low return quantiles. In other words, as returns increase at the lower quantiles of returns, investor agreement falls (investor disagreement increases). Thus, the results show that the conversations on the microblogs largely show disagreements about the prospects of the mentioned tickers at lower quantiles while at higher return quantiles, investors generally agree on the prospects of the companies mentioned on the microblogs. The high dispersion of investor opinions on the microblogs at low return quantiles corroborates the relationship between bullishness and stock returns at different return quantiles discussed in the previous section. In Section 4.3.1 bullishness is shown to be significantly associated with returns at lower return quantiles, which are the return quantiles that are also associated with increased disagreement. At higher return quantiles, bullishness is not significantly associated with returns because of the low dispersion of investor opinion and therefore less positive serial correlations in returns. Again, the OLS estimate does not capture the relationship between investor agreement and stock returns at the high and low quantiles of the latter.

4.3.3 Message volume and stock returns

The results on the magnitude of the relationship between message volume and stock returns in Table 5 show that the relationship is significant ($p < 0.01$) across all quantiles of returns. However, at low quantiles of returns ($\tau\epsilon[0.1, 0.25]$) the association is negative while at the middle ($\tau\epsilon[0.5]$) and higher ($\tau\epsilon[0.75, 0.9]$) return quantiles, the association becomes positive. The relationship between message volume stock returns across the return distribution is visualised in Figure 6.

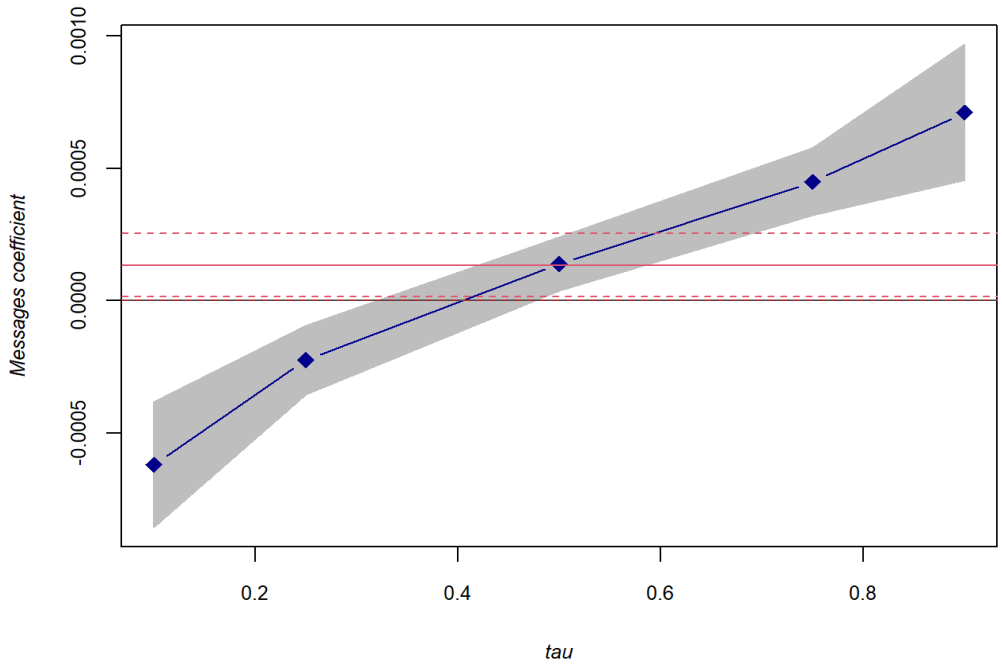


Figure 6: Message volume - QR and OLS estimates with the 95% confidence intervals

Notes: *The red continuous line shows the Ordinary Least Squares (OLS) estimate while the two dotted lines show the 95% confidence interval of the OLS estimates and the brown continuous line shows the point where the vertical axis is equal to zero. The vertical axis shows the coefficients for the message volume at the various quantiles while the horizontal axis indicates the quantile distributions of the returns* ($\tau\epsilon[0.1, 0.25, 0.5, 0.75, 0.9]$. *The values of the estimated* $\beta(\tau)$ *parameters are connected by the blue line while the grey shaded area indicates the 95% confidence intervals of the* $\beta(\tau)$ *estimated parameters.*

As Figure 6 shows, $\beta(\tau)$ increases in absolute terms as we move from the median quantile to the extreme quantiles and remains statistically significant across all the return quantiles. Like the agreement index, as the quantile levels move up, the estimate of message volume varies widely in sign, magnitude and significance. The 95% confidence intervals of the QR estimates at the very high ($\tau \in [0.75, 0.9]$) and low ($\tau \in [0.1, 0.25]$) quantiles have no overlap with the 95% confidence interval of the OLS estimate. This finding indicates that the OLS estimate does not capture the relation between message volume and stock returns at the high and low quantiles of the latter. Though the volume of messages posted reaches its maximum in the upper return quantile, this return quantile is also associated with the least value of the bullishness index. This shows a negative relationship between message volume and bullishness as we move from the lowest return quantiles to the highest return quantiles. This is consistent with the stale news hypothesis of

Tetclock (2007) which suggests that most information posted on microblogs would have already been posted and might not significantly affect sentiment since it would be stale news. High message volume may simply be indicative of a re-emphasis of what is already known and not an indication of developing information.

4.3.4 Equality of coefficients

To test if the coefficients of the *tweet* features obtained from the quantile regressions in the previous subsection are homogeneous across all the quantile levels, slope equality tests are estimated. A Wald test is estimated to determine the homogeneity of the slopes of the *tweet* features in models using returns as the dependent variables. The homogeneity test considers whether the three slope coefficients (bullishness, message volume and agreement) equal each other across the five quantiles. The null hypothesis of the Wald test is that the slope parameters are jointly equal to zero at all quantiles of the dependent variable. Specifically, the Wald test is designed to test the following null hypotheses:

$$\begin{cases} \beta k_{\tau=0.1} = \beta k_{\tau=0.25} \\ \beta k_{\tau=0.25} = \beta k_{\tau=0.5} \\ \beta k_{\tau=0.5} = \beta k_{\tau=0.75} \\ \beta k_{\tau=0.75} = \beta k_{\tau=09} \end{cases} \quad where\ k = 1,2,3,4 \qquad \text{Equation [18]}$$

Where βk_{τ} represents the coefficients for Bullishness, Agreement, Message volume and market return at $\tau \in [0.1, 0.25, 0.5, 0.75, 0.9]$. The results of the tests for the homogeneity of the *tweet* feature coefficients at all the quantile levels considered are shown in Table 6. As shown in Table 6, the overall χ^2 statistic is 1193.087 and statistically significant at the 1% significance level. Overally, the null hypothesis of equality of slopes can be rejected. This means that the coefficients of the *tweet* features are statistically different across all the return quantiles. The individual coefficient restriction tests in Table 6 are all statistically significant at the 1% level of significance except for the Bullishess score at $\tau = 0.75\ and\ \tau = 0.9$. This shows that using the individual coefficient restriction tests the null hypothesis of equality of slopes for Bullishness cannot be rejected at the top quantiles. This is corroborated in Figure 4 which suggests equality of the bullishness coefficients at the two highest quantiles.

Table 6: Quantile slope equality test

Test summary	χ^2 *statistic*	χ^2 d.f	Prob
Wald test	1193.087	16	0.0000

Restriction detail: $\beta(\tau_h)$- $\beta(k)$=0

Quantile	Variable	Restricted value	Std. Error	Prob
$\tau(0.1,0.25)$	Bullishness	0.0020	0.0006	0.0036
	Agreement	-0.0122	0.0022	0.0000
	Messages	-0.0003	0.0002	0.0001
	Market return	0.0477	0.0068	0.0000
$\tau(0.25,0.5)$	Bullishness	0.0015	0.0005	0.0036
	Agreement	-0.0062	0.0009	0.0000
	Messages	-0.0003	0.0000	0.0000
	Market return	0.1016	0.0052	0.0000
$\tau(0.5,0.75)$	Bullishness	0.0007	0.0005	0.0001
	Agreement	-0.0062	0.0009	0.0000
	Messages	-0.0003	0.0000	0.0000
	Market return	-0.1014	0.0055	0.0000
$\tau(0.75,0.9)$	Bullishness	-0.0001	0.0007	0.8071
	Agreement	-0.0134	0.0019	0.0000
	Messages	-0.0002	0.0001	0.0157
	Market return	-0.0614	0.0075	0.0000

The quantile slope equality test results point to the caveat about using a mean or median level of the data to examine the association between *tweet* features and market features. Thus, using the mean/median regression in the test equation will probably result in a loss of efficiency. This can be seen in the results in Table 4 showing an insignificant relationship between message volume and stock returns at all the conventional levels using the mean-based fixed-effects model specification. However, using quantile regression, message volume is significantly associated with stock returns at all return quantiles at the 99% confidence level.

Several points can be made from the relationship between the magnitudes of *tweet* features and market features. Firstly, investors tend to post more messages about the prospects of companies on microblogs at higher return quantiles than at lower return quantiles. However, the amount of

messages posted does not help in disaggregating the relationship between bullishness and returns. Bullishness is rather associated with returns at lower return quantiles, showing that investors put more weight on the few messages that are posted during bad times than the more messages posted during good times. Also, there is a tendency for investors to disagree (agree) a lot during bad (good) market conditions making the market more predictable during bad times. The findings from the study depart from Antweiler and Frank (2002) who reports that high message postings reflect differences in investor opinions. The findings from this study show that high messages are concentrated at higher return quantiles while at the same time the high return quantiles are associated with less investor disagreement (more agreement).

4.4 The information content of past values of *tweet* features

The contemporaneous relationships between the *tweet* features and market features discussed in Section 4.2 are crucial in comprehending whether an association between the two exist, but they do not fully portray the quality of the information of past values of firm-level *tweets* to predict future returns (Sprenger *et al*., 2014). Bollen, Mao and Zeng (2011) demonstrate that if microblogs contain new information that is not yet captured by the market prices, *tweet* features should anticipate the changes in the market features. Granger non-causality tests are appropriate for this cause as they reflect whether past values of a variable have information that can be used to forecast another variable (Bollen, Mao & Zeng, 2011). To perform the panel Granger non-causality test, the panel data should be stationary. Testing for unit roots was done for the panel data using the Levin, Lin and James Chu (2002) and the Im, Pesaran and Shin (2003) panel unit root tests as well as the ADF (Maddala & Wu, 1999) and the PP (Hadri, 2000) Fisher-type tests. Research has shown that panel unit root tests are less likely to commit a Type II error than the individual unit root tests as panel data captures more information. The results for the unit roots tests for the panel data are displayed in Table 7:

Table 7: Summary of panel unit root tests

Panel unit root summary		
Method	*Variable*	*Statistics*
Levin, Lin & Chu[a]	Bullishness	-8.6422***
Im, Pesaran and Shin[b]	Bullishness	-60.868***
ADF[c]	Bullishness	4286.4***
PP[c]	Bullishness	7459.5***
Levin, Lin & Chu	Agreement	-4.78990***
Im, Pesaran and Shin	Agreement	-46.8539***
ADF	Agreement	3586.33***
PP	Agreement	7834.67***
Levin, Lin & Chu	Messages	-24.0633***
Im, Pesaran and Shin	Messages	-64.2794***
ADF	Messages	5175.54***
PP	Messages	11427.4***
Levin, Lin & Chu	Returns	-531.086***
Im, Pesaran and Shin	Returns	-452.163***
ADF	Returns	18562.3***
PP	Returns	14024.1***

Notes: *Null hypothesis: The variable has a unit root*
****p<0.01, [a] t-test, [b]Wald statistic, [c]Fisher Chi-square*

The null hypotheses of all the unit root tests described above assume unit root processes. The results in Tables 7 show that the null hypotheses of the presence of unit roots in the bullishness, agreement, message volume and return series can be rejected at the 99% confidence level. This means that the variables are stationary ($I(0)$) processes and the estimation of the panel Granger non-causality tests will not give biased estimates. After confirming the stationarity of the variables, the Dumitrescu Hurlin Panel Causality tests are then estimated using the methodology described in Section 3.3.3. Tables 8 shows the results from the panel Granger non-causality test using one-day and two-day lags:

Table 8: Pairwise Dumitrescu Hurlin Panel Causality Tests

Pairwise Dumitrescu Hurlin Causality Tests							
Null hypothesis			*W-Stat*	*Zbar-Stat*	*p-value*	*Lags*	*Causality*
Bullishness	⇒	Returns	1.3739	0.5960	0.0653	1	NO
Bullishness	⇒	Returns	2.8990	0.5960	0.1481	2	NO
Agreement	⇒	Returns	1.5673	0.7429	0.4575	1	NO
Agreement	⇒	Returns	2.4881	0.1386	0.8897	2	NO
Messages	⇒	Returns	1.3739	0.4044	0.6859	1	NO
Messages	⇒	Returns	2.8990	0.5960	0.5512	2	NO
Returns	⇒	Bullishness	3.0226	3.2896	0.0000	1	YES
Returns	⇒	Bullishness	3.8756	1.6832	0.0000	2	YES
Returns	⇒	Agreement	3.2555	3.6971	0.0002	1	YES
Returns	⇒	Agreement	4.8243	2.7392	0.0062	2	YES
Returns	⇒	Messages	3.0226	3.2896	0.0010	1	YES
Returns	⇒	Messages	3.8756	1.6832	0.0023	2	YES

Note: *Null hypotheses: Tweet features do not homogenously cause Returns*
Returns do not homogenously cause Tweet features

In line with previous studies which set the rejection criterion at 5%, using one-day lag, the results in Table 8 leads us to fail to reject the null hypothesis that the *tweet* features do not Granger cause returns. This means that the lagged values of *tweet* features by one day do not contain important information that could be used to predict stock returns. On the other hand, we reject the null hypothesis that stock returns do not homogenously Granger cause *tweet* features at the 1% level of significance. This means that the lagged values of stock returns by one day contain useful information that can be used to predict *tweet* features in the next period. The results using two lags are also qualitatively similar, the *tweet* features at time $t-2$ do not contain statistically significant and useful information that can be used to predict the stock returns at time t. Conversely, it is rather the stock returns at time $t-2$ that contain information that can be used to predict *tweet* features at time t. These results are qualitatively similar to Kim and Kim (2014), who, using panel data and the Granger non-causality method, found no evidence that investor sentiment from internet message boards could predict stock returns either at the individual or aggregate level. Rather, the authors report that investor sentiment extracted from internet message boards through textual analysis is positively affected by previous stock performance.

4.5 Robustness Checks[10]

4.5.1 Quantile regressions using alternative conditional quantiles

The results of the quantile regression estimated in Section 4.3 are based on the conditional quantiles $\tau = \epsilon[0.1, 0.25, 0.5, 0.75, 0.9]$. To ensure the robustness of the quantile regression estimates, the model is re-estimated with alternative conditional quantiles that have also been used in previous studies. Firstly, the quantile regression model is re-estimated using $\tau = \epsilon[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ in line with Ni and Xu (2015) and then using $\tau = \epsilon[0.05, 0.2, 0.4, 0.6, 0.8, 0.95]$[11] in line with Ma and Zonggang (2018) with the statistics and visualisation of the results from the first model shown in Table 9 and Figure 7 respectively:

Table 9: Results from quantile regressions using $\tau = \epsilon[0.1:0.9]$[12]

			Independent variables					
Dependent variable	Stock return	τ	Bullishness	Agreement	Messages	Market return	Constant	Pseudo R^2
		0.1	0.005*** (0.001)	-0.019*** (0.00241)	-0.0006*** (0.00012)	0.651*** (0.0099)	-0.022*** (0.00010)	0.0426
		0.2	0.0037*** (0.001)	-0.0096*** (0.00122)	-0.0002*** (0.00008)	0.624*** (0.0067)	-0.013*** (0.00007)	0.0454
		0.3	0.0031*** (0.00050)	-0.006*** (0.00100)	-0.0002*** (0.00006)	0.579*** (0.0054)	-0.007*** (0.00006)	0.0440
		0.4	0.00255*** (0.00043)	-0.0037*** (0.00090)	-0.00002*** (0.00005)	0.52740*** (0.00443)	-0.0036*** (0.00005)	0.0440
		0.5	0.00171*** (0.00039)	-0.0014*** (0.00079)	0.00014*** (0.00005)	0.50134*** (0.00407)	-0.00035*** (0.00004)	0.0314
		0.6	0.00156*** (0.00042)	0.0004 (0.00084)	0.00027*** (0.00005)	0.52345*** (0.00407)	0.00294*** (0.00005)	0.0382
		0.7	0.00066 (0.00051)	0.00259** (0.00105)	0.00040*** (0.00006)	0.57431*** (0.00495)	0.00696*** (0.00006)	0.0423
		0.8	0.00118* (0.00065)	0.0065*** (0.00135)	0.00052*** (0.00008)	0.62584*** (0.00667)	0.01249*** (0.00007)	0.0433
		0.9	0.00115 (0.00104)	0.0183*** (0.00228)	0.00071*** (0.00013)	0.66432*** (0.01026)	0.02160*** (0.00011)	0.0380

Notes:

1. *The first row for each return quantile reports the regression coefficients for each variable, the second row reports the standard errors (in brackets)*
2. **p<0.1; **p<0.05; ***p<0.01*

[10] The R and Eviews output from the robustness tests is included as Appendix C

[11] The results for this quantile distribution are reported in Appendix C, Table C9 and Figure C1 with the results largely corroborating the previous results reported using alternative quantile distributions

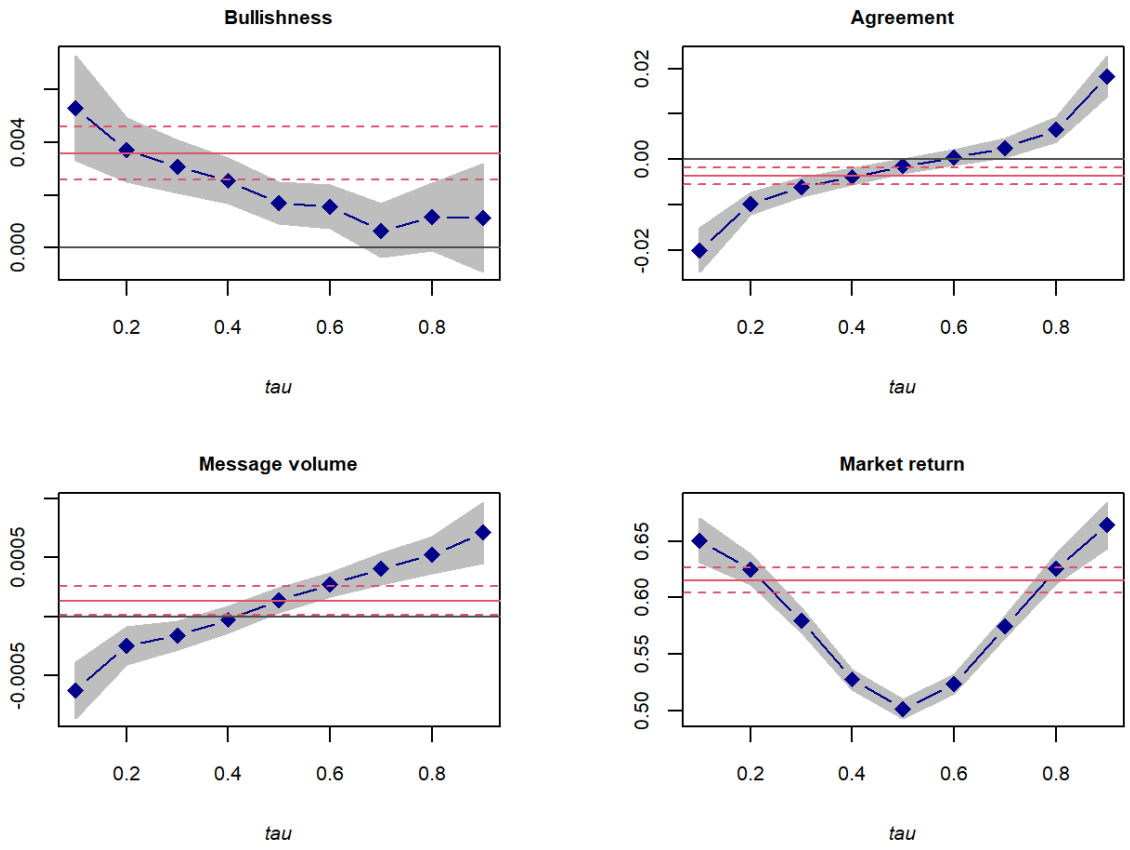


Figure 7: Visualising quantile estimators using $\tau = \epsilon[0.1{:}\,0.9]$

Notes: *The red continuous line shows the Ordinary Least Squares (OLS) estimates while the two dotted lines show the 95% confidence interval of the OLS estimates and the brown continuous line shows the point where the vertical axis is equal to zero. The vertical axes show the coefficients for the bullishness score, agreement index, message volume and market return at the various quantiles while the horizontal axes indicate the quantile distributions of the returns ($\tau\epsilon[0.1, 0.25, 0.5, 0.75, 0.9]$. The values of the estimated $\beta(\tau)$ parameters are connected by the blue line while the grey shaded area indicates the 95% confidence intervals of the $\beta(\tau)$ estimated parameters.*

As shown in Table 9 and Figure 7, the results from the quantile regression using an alternative quantile distribution are qualitatively similar to the previous results reported in Section 4.3. For investor agreement, as the quantile levels move up, the estimates of the agreement index vary widely in sign, magnitude and significance while message volume shows a negative relationship at lower return quantiles $\tau = \epsilon[0.1, 0.2, 0.3, 0.4]$ and a positive relationship at higher return quantiles $\tau = \epsilon[0.6, 0.7, 0.8, 0.9]$. Bullishness is positively related to returns at all return quantiles but the magnitude is high at lower quantiles and decreases as we move to high return quantiles until it becomes flat and insignificant at $\tau = \epsilon[0.7]$. For market return, the magnitude of the coefficients rises as we move from the middle quantile to the lower as well as upper return

quantiles respectively. Thus the use of alternative conditional quantiles does not alter the structure of the relationship between the *tweet* features and market features as well as the control variable at all return quantiles.

4.5.2 Fama and Macbeth style regressions

The findings from Section 4.4 show that using panel Granger non-causality tests, *tweet* features lagged by one day and two days, do not contain information that can be used to forecast stock returns. Instead of using the panel Granger non-causality methodology to examine the information content of *tweet*s, some studies (such as. Loughran & Mcdonald, 2011) have also utilised the lead-lag Fama-Macbeth style regressions. This methodology is also utilised to test the robustness of the panel Granger non-causality tests reported in Section 4.4. Firstly, the stock returns are regressed on the 1-day and 2-day lagged values of bullishness, agreement and message volume. Conversely, *tweet* features are regressed on the 1-day and 2-day lagged values of stock returns to see if past values of stock returns contain information that can be useful in predicting future values of *tweet* features. The results of the Fama and Macbeth regressions of stock returns on lagged values of *tweet* features are reported first in Table 10:

Table 10: Lead-lag Fama Macbeth style regressions of returns on tweet features

	Dependent Variable		
	Stock Returns		
	(1)	*(2)*	*(3)*
Bullishness_{t-1}	-0.005		
	(0.007)		
Bullishness_{t-2}	0.005		
	(0.007)		
Agreement_{t-1}		-0.002	
		(0.003)	
Agreement_{t-2}		0.000	
		(0.000)	
Messages_{t-1}			-0.000
			(0.000)
Messages_{t-2}			0.000
			(0.000)
Size	0.0002***	0.0002***	0.0002***
	(0.002)	(0.002)	(0.002)
Constant	-0.005***	-0.005***	-0.006***
	(0.002)	(0.002)	(0.002)
Observations	175 429	175 429	175 429
R^2	0.117	0.125	0.130
Notes:	**p<0.1; **p<0.05; ***p<0.01*		

Table 10 shows the results from the 3 models used to test whether past values of *tweet* features have information content that can be used to predict future values of stock returns. Models 1, 2 and 3 are the regressions of stock returns on the lagged values of bullishness, investor agreement and message volume respectively. In all the model specifications, the market capitalisation of the companies is used as a control variable in line with previous studies. The coefficients of the lagged *tweets* features are all statistically insignificant at the conventional significance levels. Consistent with Sprenger *et al*., (2014), these results provide evidence that, though *tweet* features are contemporaneously associated with market features, past values do not contain statistically significant information which can be used to predict future stock returns. Table 11 shows the results on the regressions of the *tweet* features (bullishness, messages, agreement) on the 1-day and 2-day lagged values of stock returns.

Table 11: Lead-lag Fama Macbeth style regressions of returns on tweet features

	Dependent variable		
		Tweet Features	
	Bullishness	Messages	Agreement
$Return_{t-1}$	0.115***	-0.327***	-0.031***
	(0.017)	(0.148)	(0.010)
$Return_{t-2}$	0.071***	-0.273*	-0.031***
	(0.016)	(0.142)	(0.010)
Size	0002***	0.278***	0.010***
	(0.0002)	(0.002)	(0.0001)
Constant	-0.048***	-6.233***	-0.214***
	(0.006)	(0.056)	(0.003)
Observations	175 253	175 253	175 253
R^2	0.021	0.222	0.086

Notes: **p<0.1, **p<0.05, ***p<0.01*

As shown in Table 11, the 1-day and 2-day lagged coefficients of returns are all statistically significant at the 5% level except the 2-day lag coefficient of returns in the model using message volume as the dependent variable. This means that past values of stock returns have useful information that can be used to predict the future values of the *tweet* features. Higher past returns are likely to predict increased investor bullishness while higher past returns are likely to signal fewer messages being posted on stock microblogs. The relationship between investor agreement and lagged values of stock returns shows that low past returns are likely to signal less investor opinion dispersion in the future. Not surprisingly, the coefficients of the one day-lagged returns are higher than those of the two days-lagged returns showing that one-day lagged returns contain more information than two-days lagged returns.

4.5.3 Granger non-causality test using weekly and monthly data

The predictive power of *tweet* features on stock returns is further investigated using weekly and monthly data. Various studies that have examined the information content of investor sentiment in a South African context have mostly used weekly and monthly survey data and the majority have confirmed the information content of investor sentiment in predicting future values of stock returns (such as Dalika & Seetharam, 2015; Rupande, Muguto & Muzindutsi, 2019). To find the weekly and monthly scores for message volume and investor agreement, the daily *tweet* features are aggregated at the end of every week and month respectively, with the scores calculated as

previously specified in Section 3.2.1. Following Kim and Kim (2014), the weekly and monthly scores for bullishness are calculated by getting the weekly and monthly aggregates of all messages with positive sentiment and negative sentiment ($Pos_{i,t}$ and $Neg_{i,t}$ respectively) for firm i at time t. The weekly and monthly bullishness scores are then calculated in line with Antweiler and Frank (2004) using the following formula:

$$B_{i,t} = ln\left(\frac{1 + Pos_{i,t}}{1 + Neg_{i,t}}\right) \qquad \text{Equation [19]}$$

Where $B_{i,t}$ is the bullishness score for firm i at time t, where time is at weekly and monthly intervals. The weekly and monthly returns are calculated as the natural logarithm of the stock price at $week\ t$ ($month\ t$) divided by the stock price at $week\ t-1$ ($month\ t-1$). Granger non-causality is estimated using a simple F-test where the null hypothesis is that the *tweet* features do not Granger cause stock returns and *vice-versa.* Before estimating the Granger non-causality tests, the weekly and monthly data are first tested for stationarity using the Levin, Lin and James Chu (2002) and the Im, Pesaran and Shin (2003) panel unit root tests as well as the ADF (Maddala & Wu, 1999) and the PP (Hadri, 2000) Fisher-type tests. The results from the unit root tests are shown in Table 12 below:

Table 12: Panel unit root tests using weekly and monthly data

Panel Unit Root Tests Summary			
Method	*Variable*	*Statistic*	
		Weekly Data	*Monthly Data*
Levin, Lin & Chu[a]	Bullishness	-4.32415***	-115.803***
Im, Pesaran and Shin[b]	Bullishness	-20.8024***	-28.3135***
ADF[c]	Bullishness	1584.79***	513.070***
PP[c]	Bullishness	1950.38***	607.004***
Levin, Lin & Chu	Agreement	-5.91274***	-12.8878***
Im, Pesaran and Shin	Agreement	-29.5134***	-17.7793***
ADF	Agreement	2297.93***	920.821***
PP	Agreement	2574.64***	1052.23***
Levin, Lin & Chu	Messages	-36.1971***	-18.7388***
Im, Pesaran and Shin	Messages	-56.9962***	-26.0641***
ADF	Messages	4443.82***	1277.03***
PP	Messages	6391.30***	1408.00***
Levin, Lin & Chu[a]	Returns	-200.564***	-3.97622***
Im, Pesaran and Shin	Returns	-189.144***	-4.20289***
ADF	Returns	16710.5***	390.008***
PP	Returns	17392.1***	760.627***

Notes: *Null hypothesis: The variable has a unit root*
****p<0.01, [a] t-test, [b]Wald statistic, [c]Fisher chi-square*

Table 12 shows that the null hypotheses of unit roots can be rejected for all the variables at the 99% confidence intervals. This means that the variables are stationary $(I(0))$ processes and therefore Granger non-causality tests would not give biased estimates. Table 13 shows the results from the panel Granger non-causality F-tests using weekly and monthly data as described above.

Table 13: Pairwise Granger causality tests using weekly data

Pairwise Granger Causality Tests (Weekly Data)							
Null Hypothesis			*Observations*	*F Statistic*	*p-value*	*Lags*	*Causality*
Bullishness	⇒	Returns	34 045	0.99394	0.3188	1	YES
Bullishness	⇒	Returns	32 874	1.93459	0.1445	2	YES
Agreement	⇒	Returns	34 045	0.97085	0.3245	1	YES
Agreement	⇒	Returns	32 874	2.98096	0.0508	2	YES
Messages	⇒	Returns	34 045	0.01522	0.9018	1	YES
Messages	⇒	Returns	32 874	0.42453	0.6541	2	YES
Returns	⇒	Bullishness	34 045	1.04690	0.3062	1	YES
Returns	⇒	Bullishness	32 874	2.63324	0.0719	2	YES
Returns	⇒	Agreement	34 045	0.00235	0.9614	1	YES
Returns	⇒	Agreement	32 874	0.52005	0.5945	2	YES
Returns	⇒	Messages	34 045	0.52626	0.4682	1	YES
Returns	⇒	Messages	32 874	0.82600	0.4378	2	YES

The results in Table 13 show that using weekly data, the null hypotheses that *tweet* features do not Granger cause stock returns can be rejected using one week and two-week lags. Thus, the results show that weekly *tweet* features contain useful information that could be used to predict weekly future stock returns. The results in Table 13 also attest to the presence of bidirectional causality between *tweet* features and stock returns as it can be seen that weekly stock returns also contain important information that could be used to predict future weekly *tweet* features. The results on the causal relationship between *tweet* features and stock returns are also similar using monthly data as shown in Table 14:

Table 14: Pairwise Granger causality tests using monthly data

Pairwise Granger Causality Tests (Monthly Data)							
Null Hypothesis			*Observations*	*F Statistic*	*p-value*	*Lags*	*Causality*
Bullishness	⇒	Returns	3 238	0.25854	0.6112	1	YES
Bullishness	⇒	Returns	2 915	0.21604	0.8057	2	YES
Agreement	⇒	Returns	3 238	0.21646	0.6418	1	YES
Agreement	⇒	Returns	2 915	0.04940	0.9518	2	YES
Messages	⇒	Returns	3 238	1.12450	0.2890	1	YES
Messages	⇒	Returns	2 915	3.07479	0.0563	2	YES
Returns	⇒	Bullishness	3 238	2.5E-06	0.9987	1	YES
Returns	⇒	Bullishness	2 915	0.02114	0.9791	2	YES
Returns	⇒	Agreement	3 238	0.00014	0.9907	1	YES
Returns	⇒	Agreement	2 915	0.60647	0.5453	2	YES
Returns	⇒	Messages	3 238	0.81981	0.3653	1	YES
Returns	⇒	Messages	2 915	0.27405	0.7603	2	YES

Past monthly metrics of *tweet* features have information content that could be used to predict future stock returns using one-month and two-month lags. It can be seen that *tweet* features have no predictive power on stock returns when using high-frequency daily granularity analysis. At lower weekly and monthly granularity analysis, *tweet* features are found to be useful in predicting future stock returns. These results suggest that stock returns are random at daily intervals while at the lower weekly and monthly intervals stock returns can be predicted. These results provide evidence that might point to cyclical market efficiency on the Johannesburg Stock Exchange where the stock market goes through intermittent periods of efficiency and inefficiency. The results also corroborate Seetharam, Auret and Celik (2017) who document evidence of the dynamic nature of market efficiency on the JSE. This points to the Adaptive Market Hypothesis as a potential explanation of the nature of efficiency of the JSE.

4.5.4 Brock-Dechert-Scheinkman (BDS) tests for non-linearity

An important result from Section 4.3 is that the relationships between stock returns and *tweet* features are not monotonic across the distribution of the returns. This could prove that the relationship between stock returns and *tweet* features is nonlinear and should not, therefore, be modelled using linear specifications. To confirm whether the relationship between stock returns and *tweet* features is linear or not, the Brock-Dechert-Scheinkman (BDS) is applied on the residuals from the stock return regression on each of the *tweet* features. The BDS statistic tests the

null that the series in question are independent and identically distributed (*i.i.d*) against an unspecified alternative using a non-parametric technique, and has power against a variety of nonlinear processes. The process is done for each of the companies that formed part of the sample used in this study. For brevity, only the results of a few selected counters are reported but the results for all the companies are qualitatively similar. The results of the BDS tests for selected companies are shown in Table 15 below.

Table 15: BDS statistics on residuals from the regression of returns on tweet features

Firm	Independent Variable	*m*				
		2	3	4	5	6
ABG	Bullishness	3.2872***	4.8562***	5.2971***	5.6585***	5.7927***
	Messages	3.2802***	4.8662***	5.3169***	5.6906***	5.8362***
	Agreement	3.2473***	4.7999***	5.2008***	5.5487***	5.6690***
ACL	Bullishness	5.6289***	7.8611***	8.4502***	9.1931***	9.8502***
	Messages	5.6534***	7.9154***	8.5321***	9.2966***	9.9631***
	Agreement	5.7002***	7.9545***	8.5529***	9.3274***	10.007***
ADH	Bullishness	8.5130***	9.9336***	10.134***	10.432***	10.715***
	Messages	8.5447***	9.9773***	10.159***	10.457***	10.740***
	Agreement	8.5232***	9.9457***	10.146***	10.448***	10.729***

Notes: *Entries correspond to the z-statistic of the BDS test with the null of i.i.d. residuals, with the test applied to the residuals recovered from the stock return equation on the tweet features; m denotes the embedding dimension (the number of consecutive data points to include in the set), *** indicates rejection of the null hypothesis at 1 percent level of significance. All the stocks used in the study were individually tested but for brevity, only selected stocks are reported. However, the results for all the stocks are qualitatively similar.*

The results above give evidence of the rejection of *i.i.d* residuals at different embedded dimensions at the highest level of statistical significance. This proves that there are nonlinear relationships between stock returns and *tweet* features. While the results from the BDS test show that there are nonlinear relationships between stock returns and *tweet* features, research has shown that most of the non-linear dependence in stock returns exist as a result of neglecting conditional heteroscedasticity that could be captured through ARCH/GARCH models. The BDS test has been found to have high power against ARCH/GARCH where nonlinearity has entered through the conditional variance. A GARCH (1,1) is therefore fitted on the returns and the resulting standardised residuals are then tested for *i.i.d* using the BDS test. A further rejection of *i.i.d* would

suggest that conditional heteroscedasticity is not the main source of nonlinearity. The BDS test is conducted on the residuals recovered from a GARCH (1,1) process fitted on the return series of all the firms that formed part of the sample of this study and, for all the return series, the null hypothesis of linearity is rejected showing that the nonlinear dependence is of an unknown form in the data after volatility clustering is removed. The results for selected companies are shown in the table below:

Table 16*: *BDS statistics on residuals recovered from GARCH (1,1) fitted on stock returns

Firm	Variable	*m*				
		2	3	4	5	6
ABG	Stock return	3.2904***	4.8597***	5.2975***	5.6583***	5.7936***
ACL	Stock return	5.6409***	7.8914***	8.4963***	9.2577***	9.9296***
ADH	Stock return	8.5351***	9.9539***	10.147***	10.445***	10.727***
AEL	Stock return	3.4884***	4.1547***	4.6386***	4.5764***	4.5764***

Notes: *Entries correspond to the z-statistic of the BDS test with the null of i.i.d. residuals, with the test applied to the residuals recovered from the GARCH (1,1) fitted on the stock returns, m denotes the embedding dimension (the number of consecutive data points to include in the set), *** indicates rejection of the null hypothesis at 1 percent level of significance. All the stocks used in the study were individually tested but for brevity, only selected stocks are reported. However, the results for all the stocks are qualitatively similar.*

As shown in Table 16, the null hypothesis of *i.i.d* residuals can be rejected at the 1% significance level at various embedding dimensions (*m*) ranging from 2 to 6. The results, therefore, show that the nonlinear dependence of the return series of the stocks used in the study is of an unknown form after volatility clustering is removed. This indicates the necessity for employing nonlinear specifications for modelling the relationships between investor sentiment and stock returns.

5 CONCLUSION

The study purposed to answer three research questions namely; whether there is a contemporaneous link between *tweet* features and stock market features; whether the magnitude of *tweet* features is monotonically related to the magnitude of market features; and whether past values of *tweet* features have information content that could be used to forecast the future values of stock returns. Using the fixed-effects model specification consistent with the model selected using the Hausman test, the study finds that except for the relationship between message volume and stock returns, and between bullishness and stock returns, *tweet* features are contemporaneously associated with stock returns.

While the results above largely show contemporaneous associations between *tweet* features and market features at the mean, they do not give a full picture of whether the magnitude of the relationship between the two features is monotonic. On the link between the magnitude of *tweet* features and stock market features, the findings from the pooled quantile regression and quantile regression with firm fixed effects show that there is no monotonic link between the magnitude of *tweet* features and market features. The results reveal a monotonic relationship between the magnitude of bullishness and stock returns at low and middle quantile of returns while at high return quantiles the relationship becomes flat and statistically insignificant. On the other hand, message volume and investor agreement have stronger relationships with stock returns in absolute terms at extreme return quantiles. The results show no evidence of a monotonic link between the magnitude of *tweet* features and market features as there are asymmetric spillover effects of *tweet* features on the stock market returns.

One of the most crucial questions that researchers have grappled with in recent times is whether online messages contain useful content that can be used to predict asset prices. Using panel Granger non-causality tests and lead-lag Fama-Macbeth style regressions, the results from this study show no evidence of past values containing important information that could forecast stock market returns in the next period using daily data. The evidence rather shows that stock returns contain useful information that could be used to predict stock market returns in the next period. However, utilising weekly and monthly data, there is strong evidence that historical values of *tweet*

features can be used to predict future returns. The non-predictability of stock market returns from *tweet* features using daily *tweet* and return data and predictability using weekly and monthly data confirm the results of Seetharam, Auret and Celik (2017). The trio report that the daily return behaviour of the JSE All Share Index follows a random walk while at lower frequencies the return series exhibit non-randomness.

This research report examined the link between textual sentiment mined from stock microblogs (Twitter and StockTwits) and market features. However, these market features are only available as aggregate market indicators. Further studies could replace trading volume with the number of trades of different size categories to distinguish between institutional and retail investors since existing literature shows that textual sentiment is mainly associated with individual investors compared to smart investors. Additionally, future research could examine whether the contemporaneous and intertemporal relationships between *tweet* features and market features are the same in the presence of popular market anomalies.

The popularity of stock microblogs like Twitter and StockTwits, like any microblog, has been driven by the real-time nature of the platforms which allows individuals to have access to information instantaneously. This means that market features are likely to adjust to new information more rapidly than from traditional news platforms. The daily intervals used in this study, where *tweet* features are computed based on *tweets* posted throughout the day, may produce biased estimates. Future research could make use of higher-frequency data, for example using 15-minute granularity like in Li, van Dalen and van Rees (2018). Though a study of that nature may be restricted by limited message volume, it may be worthwhile to choose a few stocks with adequate message volume for this purpose. Since the BDS test provided evidence of the nonlinear relationships between stock returns and tweet features, future studies could also unravel the relationship using non-parametric models based on nonlinear dependence.

Finally, though the emergence of social media platforms has provided investors with an alternative source of information, oftentimes the large amounts of information from these platforms can contain potential noise that can mislead investors. This is especially true on the back of the increase in bots, cybots and social media farms (Ferrara *et al*., 2016). These social bots could be used to

intentionally manipulate the stock market by disseminating news towards a predetermined objective. Future studies could look at the impact of bot-created *tweets* on the stock market to establish if the impact is different from human-created *tweets.*

In conclusion, the study has established no causal effects between the *tweet* features and the market features using daily granularity. This confirms and implies that the JSE is being dominated by institutional investors rather than noise traders who normally trade on sentiment. Secondly, the lack of causal effects between *tweet* features and market features at the daily granularity and the existence of the relationship using weekly and monthly data confirms the dynamic nature of the efficiency of the JSE. This implies that policymakers have to implement appropriate regulations to deter the development of bubbles or crashes during the *"greed"* and *"fear"* cycles. For asset allocation purposes the findings from this study imply that for a better way of achieving consistent levels of expected returns, active allocation strategies should be dynamic and respond to, and adapt to changing market conditions.

REFERENCES

Agrawal, S., Azar, P.D., Lo, A.W. & Singh, T. 2018. Momentum, Mean-Reversion, and Social Media: Evidence from StockTwits and Twitter. *Journal of Portfolio Management*. 44(7):85–95.

Allen, D.E., McAleer, M. & Singh, A.K. 2019. Daily Market News Sentiment and Stock Prices. *Applied Economics*. 51(30):3212–3235.

Al-Nasseri, A. & Menla Ali, F. 2018. What Does Investors' Online Divergence of Opinion Tell us About Stock Returns and Trading Volume? *Journal of Business Research*. 86:166–178. DOI: 10.1016/j.jbusres.2018.01.006.

Antweiler, W. & 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. DOI: 10.1111/j.1540-6261.2004.00662.x.

Bachelier, L. 1901. Théorie mathématique du jeu. *Annales scientifiques de l'École normale supérieure*. 18:143–209. DOI: 10.24033/asens.493.

Baker, M. & Wurgler, J. 2006. Investor Sentiment and the Cross-Section of Stock Returns. *Journal of Finance*. 61(4):1645–1680.

Barberis, N. & Thaler, R. 2003. A survey of Behavioural Finance. In *Handbook of the Economics of Finance*.20. Elsevier. 1053–1128.

Barberis, N., Shleifer, A. & Vishny, R. 1998. A Model of Investor Sentiment. *Journal of Financial Economics*. 49(3):307–343.

Bartov, E., Faurel, L. & Mohanram, P.S. 2018. Can Twitter Help Predict Firm-Level Earnings and Stock Returns? *The Accounting Review*. 93(3):25–57. DOI: 10.2308/accr-51865.

Black, F. 1986. Noise. *Journal of Finance*. 41(3):529–43.

Bollen, J., Mao, H. & Zeng, X.-J. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*. 2(1):1–8. DOI: 10.1016/j.jocs.2010.12.007.

Breitmayer, B., Massari, F. & Pelster, M. 2019. Swarm Intelligence? Stock Opinions of the Crowd and Stock Returns. *International Review of Economics & Finance*. 64:443–464. DOI: 10.1016/j.iref.2019.08.006.

Brooks, C. 2019. *Introductory Econometrics for Finance*. 4th ed. Cambridge: Cambridge University Press. DOI: 10.1017/9781108524872.

Carvalho, C., Klagge, N. & Moench, E. 2011. The Persistent Effects of a False News Shock. *Journal of Empirical Finance*. 18:597–615. DOI: 10.2139/ssrn.1363762.

Chakraborty, M. & Subramaniam, S. 2020. Asymmetric Relationship of Investor Sentiment with Stock Return and Volatility: Evidence from India. *Review of Behavioral Finance*. (ahead-of-print). DOI: 10.1108/RBF-07-2019-0094.

Chau, F., Deesomsak, R. & Koutmos, D. 2016. Does Investor Sentiment Really Matter? *International Review of Financial Analysis*. 48:221–232. DOI: 10.1016/j.irfa.2016.10.003.

Chou, R.Y., Chou, H. & Liu, N. 2015. Range Volatility: A Review of Models and Empirical Studies. In *Handbook of Financial Econometrics and Statistics*. C.F. Lee & J.C. Lee, Eds. New York, NY: Springer. 2029–2050. DOI: 10.1007/978-1-4614-7750-1_74.

Clement, J. 2020. *Number of Social Media Users Worldwide*. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ [2020, July 24].

Cole, B. & Daigle, J. 2015. Do Tweets Matter for Shareholders? An Empirical Analysis. *Journal of Accounting and Finance*. 15(3):39–52.

Cowles, A. 1933. Can Stock Market Forecasters Forecast? *Econometrica*. 1(3):309. DOI: 10.2307/1907042.

Dalika, N. & Seetharam, Y. 2015. Sentiment and Returns: An Analysis of Investor Sentiment in the South African Market. *Investment Management and Financial Innovations*. 12(1):267–276.

Daniel, K., Hirshleifer, D. & Subrahmanyam, A. 1998. Investor Psychology and Security Market Under- and Overreactions. *The Journal of Finance*. 53(6):1839–1885. DOI: 10.1111/0022-1082.00077.

Das, S., Martínez-Jerez, A. & Tufano, P. 2005. eInformation: A Clinical Study of Investor Discussion and Sentiment. *Financial Management*. 34(3):103–137.

Daszynska-Zygadlo, K., Szpulak, A. & Szyszka, A. 2014. Investor Sentiment, Optimism and Excess Stock Market Returns. Evidence from Emerging Markets. *Business and Economic Horizons*. 10(4):362–373. DOI: 10.15208/beh.2014.27.

De Long, J.B., Shleifer, A., Summers, L.H. & Waldmann, R.J. 1990. Noise Trader Risk in Financial Markets. *Journal of Political Economy*. 98(4):703–738.

Demers, E. & Vega, C. 2008. Soft Information in Earnings Announcements: News or Noise? *International Finance Discussion Papers 951*. Board of Governors of the Federal Reserve System (U.S.).

Diether, K.B., Malloy, C.J. & Scherbina, A. 2002. Differences of Opinion and the Cross-Section of Stock Returns. *The Journal of Finance*. 57(5):2113–2141.

Dumitrescu, E.I. & Hurlin, C. 2012. Testing for Granger Non-causality in Heterogeneous Panels. *Economic Modelling*. 29(4):1450–1460. DOI: 10.1016/j.econmod.2012.02.014.

Fama, E.F. 1965. The Behavior of Stock-Market Prices. *The Journal of Business*. 38(1):34–105.

Fama, E.F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*. 25(2):383–417. DOI: 10.2307/2325486.

Fama, E.F. 1991. Efficient Capital Markets: II. *The Journal of Finance*. 46(5):1575–1617. DOI: 10.1111/j.1540-6261.1991.tb04636.x.

Fama, E.F. & MacBeth, J.D. 1973. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*. 81(3):607–636. DOI: 10.1086/260061.

Ferrara, E., Varol, O., Davis, C.B., Menczer, F. & Flammini, A. 2016. *The Rise of Social Bots*. (SSRN Scholarly Paper ID 2982515). Rochester, NY: Social Science Research Network. Available: https://papers.ssrn.com/abstract=2982515 [2020, October 01].

Ferris, S.P., Hao, Q. & Liao, M.-Y. 2013. The Effect of Issuer Conservatism on IPO Pricing and Performance. *Review of Finance*. 17(3):993–1027. DOI: 10.1093/rof/rfs018.

Galton, F. 1907. Vox Populi. *Nature*. 75(1949):450–451. DOI: 10.1038/075450a0.

García, D. 2013. Sentiment During Recessions. *The Journal of Finance*. 68(3):1267–1300. DOI: 10.1111/jofi.12027.

Giannini, R., Irvine, P. & Shu, T. 2019. The Convergence and Divergence of Investors' Opinions Around Earnings News: Evidence from a Social Network. *Journal of Financial Markets*. 42(C):94–120.

Gino, F., Brooks, A.W. & Schweitzer, M.E. 2012. Anxiety, Advice, and the Ability to Discern: Feeling Anxious Motivates Individuals to Seek and use Advice. *Journal of Personality and Social Psychology*. 102(3):497–512. DOI: 10.1037/a0026413.

Granger, C.W.J. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*. 37(3):424–438. DOI: 10.2307/1912791.

Hadri, K. 2000. Testing for Stationarity in Heterogeneous Panel Data. *The Econometrics Journal*. 3(2):148–161. DOI: 10.1111/1368-423X.00043.

Harris, M. & Raviv, A. 1993. Differences of Opinion Make a Horse Race. *The Review of Financial Studies*. 6(3):473–506. DOI: 10.1093/rfs/5.3.473.

Henry, E. & Leone, A.J. 2016. Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. *The Accounting Review*. 91(1):153–178. DOI: https://doi.org/10.2308/accr-51161.

Heston, S.L. & Sinha, N.R. 2017. News vs. Sentiment: Predicting Stock Returns from News Stories. *Financial Analysts Journal*. 73(3):67–83. DOI: 10.2469/faj.v73.n3.3.

Higgins, T. 2020. Elon Musk Tweeted That Tesla's Stock Was Too High. The Market Agreed. *Wall Street Journal*. 1 May. Available: https://www.wsj.com/articles/tesla-stock-falls-after-ceo-tweets-stock-is-too-high-11588348672 [2020, May 20].

Hillert, A., Jacobs, H. & Müller, S. 2018. Journalist Disagreement. *Journal of Financial Markets*. 41:57–76. DOI: 10.1016/j.finmar.2018.09.002.

Hirshleifer, D. & Teoh, S.H. 2009. Chapter 1 - Thought and Behavior Contagion in Capital Markets. In *Handbook of Financial Markets: Dynamics and Evolution*. T. Hens & K.R. Schenk-Hoppé, Eds. (Handbooks in Finance). San Diego: North-Holland. 1–56. DOI: 10.1016/B978-012374258-2.50005-1.

Huang, C., Yang, X., Yang, X. & Sheng, H. 2014. An Empirical Study of the Effect of Investor Sentiment on Returns of Different Industries. *Mathematical Problems in Engineering*. 2014:1–11. DOI: 10.1155/2014/545723.

Huberman, G. & Regev, T. 2001. Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar. *The Journal of Finance*. 56(1):387–396. DOI: 10.1111/0022-1082.00330.

Im, K.S., Pesaran, M. & Shin, Y. 2003. Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics*. 115(1):53–74.

Indaco, A. 2020. From Twitter to GDP: Estimating GDP from Social Media. *Regional Science and Urban Economics*. 85:87–96. DOI: https://doi.org/10.1016/j.regsciurbeco.2020.103591.

Jegadeesh, N. & Wu, A. 2013. Word Power: A New Approach for Content Analysis. *Journal of Financial Economics*. 110. DOI: 10.2139/ssrn.1787273.

Johnston, K. & Maree, S. 2015. Critical Insights into the Design of Big Data Analytics Research: How Twitter "moods" Predict Stock Exchange Index Movement. *South African Journal of Information and Communication*. (15):53–67. DOI: 10.23962/10539/20330.

Kahneman, D. 2003. A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist*. 58(9):697–720. DOI: 10.1037/0003-066X.58.9.697.

Kahneman, D. & Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*. 47(2):263–91.

Karpoff, J.M. 1986. A Theory of Trading Volume. *The Journal of Finance*. 41(5):1069–1087. DOI: 10.1111/j.1540-6261.1986.tb02531.x.

Kearney, C. & Liu, S. 2014. Textual Sentiment in Finance: A Survey of Methods and Models. *International Review of Financial Analysis*. 33:171–185. DOI: 10.1016/j.irfa.2014.02.006.

Kim, O. & Verrecchia, R.E. 1991. Trading Volume and Price Reactions to Public Announcements. *Journal of Accounting Research*. 29(2):302–321. DOI: 10.2307/2491051.

Kim, S.H. & Kim, D. 2014. Investor Sentiment from Internet Message Postings and the Predictability of Stock Returns. *Journal of Economic Behavior & Organization*. 107:708–729. DOI: 10.1016/j.jebo.2014.04.015.

Koenker, R. 2004. Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis*. 91(1):74–89. DOI: 10.1016/j.jmva.2004.05.006.

Koenker, R. & Machado, J.A.F. 1999. Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*. 94(448):1296–1310. DOI: 10.1080/01621459.1999.10473882.

Koenker, R.W. & Bassett, G. 1978. Regression Quantiles. *Econometrica*. 46(1):33–50.

Lachanski, M. & Pav, S. 2017. Shy of the Character Limit: "Twitter Mood Predicts the Stock Market" Revisited. *Econ Journal Watch.*14(3):45.

Lemmon, M. & Portniaguina, E. 2006. Consumer Confidence and Asset Prices: Some Empirical Evidence. *The Review of Financial Studies*. 19(4):1499–1529. DOI: 10.1093/rfs/hhj038.

Levin, A., Lin, C.F. & James Chu, C.S. 2002. Unit Root Tests in Panel Data: Asymptotic and Finite-sample Properties. *Journal of Econometrics*. 108(1):1–24.

Li, M.Y. & Wu, J.S. 2014. Analysts' Forecast Dispersion and Stock Returns: A Quantile Regression Approach. *Journal of Behavioral Finance*. 15(3):175–183. DOI: 10.1080/15427560.2014.942420.

Li, T., van Dalen, J. & van Rees, P.J. 2018. More than Just Noise? Examining the Information Content of Stock Microblogs on Financial Markets. *Journal of Information Technology*. 33(1):50–69. DOI: 10.1057/s41265-016-0034-2.

Li, X., Shen, D. & Zhang, W. 2019. How Does Foreign Sentiment Affect the Chinese Stock Markets? Some Empirical Evidence. *China Accounting and Finance Review*. 20(3):25.

Loughran, T. & Mcdonald, B. 2011. When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*. 66(1):35–65. DOI: 10.1111/j.1540-6261.2010.01625.x.

Ma, C., Xiao, S. & Ma, Z. 2018. Investor Sentiment and the Prediction of Stock Returns: A Quantile Regression Approach. *Applied Economics*. 50(50):5401–5415. DOI: 10.1080/00036846.2018.1486993.

Maddala, G.S. & Wu, S. 1999. A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test. *Oxford Bulletin of Economics and Statistics*. 61(S1):631–652. DOI: 10.1111/1468-0084.0610s1631.

Malkiel, B.G. 2003. The Efficient Market Hypothesis and its Critics. *Journal of Economic Perspectives*. 17(1):59–82. DOI: 10.1257/089533003321164958.

Matthews, C. 2013. How Does One Fake Tweet Cause a Stock Market Crash? *Time*. (April, 24). Available: https://business.time.com/2013/04/24/how-does-one-fake-tweet-cause-a-stock-market-crash/ [2020, April 23].

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F.Å. & Lanamäki, A. 2015. The Sum of all Human Knowledge": A Systematic Review of Scholarly Research on the Content of Wikipedia. *Journal of the Association for Information Science and Technology*. 66(2):219–245. DOI: 10.1002/asi.23172.

Milgrom, P. & Stokey, N. 1982. Information, Trade and Common Knowledge. *Journal of Economic Theory*. 26(1):17–27. DOI: 10.1016/0022-0531(82)90046-1.

Moss, A., Naughton, J. & Wang, C. 2020. The Irrelevance of ESG Disclosure to Retail Investors: Evidence from Robinhood. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3604847.

Muguto, H.T. & Rupande, L. & Muzindutsi, P.F. 2019. Investor Sentiment and Foreign Financial Flows: Evidence from South Africa. *Zb. rad. Ekon. fak. Rij.* 37(2):27.

Ni, Z.X., Wang, D.Z. & Xue, W.J. 2015. Investor Sentiment and its Nonlinear Effect on Stock returns: New Evidence from the Chinese Stock Market Based on Panel Quantile Regression Model. *Economic Modelling*. 50:266–274. DOI: 10.1016/j.econmod.2015.07.007.

Parkinson, M. 1980. The Extreme Value Method for Estimating the Variance of the Rate of Return. *The Journal of Business*. 53(1):61. DOI: 10.1086/296071.

Peterson, R.L. 2016. *Trading on Sentiment: The Power of Minds Over Markets*. John Wiley & Sons.

Price, S.M., Doran, J.S., Peterson, D.R. & Bliss, B.A. 2012. Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *Journal of Banking & Finance*. 36(4):992–1011. DOI: 10.1016/j.jbankfin.2011.10.013.

Regnault, J. 1863. *Calcul des chances et philosophie de la bourse*. Paris : Mallet-Bachelier [et] Castel. Available: http://archive.org/details/calculdeschances00regn [2020, June 09].

Rupande, L., Muguto, H.T. & Muzindutsi, P.F. 2019. Investor Sentiment and Stock Return Volatility: Evidence from the Johannesburg Stock Exchange. *Cogent Economics & Finance*. 7(1). DOI: 10.1080/23322039.2019.1600233.

Seetharam, Y. 2021. Investigating the Low-risk Anomaly in South Africa. *Review of Behavioral Finance*. (ahead-of-print). DOI: 10.1108/RBF-07-2020-0167.

Seetharam, Y., Auret, C. & Celik, T. 2017. The Dynamics of Market Efficiency: Testing the Random Walk Hypothesis in South Africa. *Frontiers in Finance and Economics.*14(1):29-69.

Shiller, R. 1981. Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *American Economic Review*. 71(3):421–36.

Shiller, R.J. 2003. From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives*. 17(1):83–104. DOI: 10.1257/089533003321164967.

Shleifer, A. & Summers, L.H. 1990. The Noise Trader Approach to Finance. *Journal of Economic Perspectives*. 4(2):19–33. DOI: 10.1257/jep.4.2.19.

Shleifer, A. & Vishny, R.W. 1997. The Limits of Arbitrage. *The Journal of Finance*. 52(1):35–55. DOI: 10.1111/j.1540-6261.1997.tb03807.x.

Sprenger, T.O., Tumasjan, A., Sandner, P.G. & Welpe, I.M. 2014. Tweets and Trades: The Information Content of Stock Microblogs: Tweets and Trades. *European Financial Management*. 20(5):926–957. DOI: 10.1111/j.1468-036X.2013.12007.x.

Stambaugh, R., Yu, J. & Yuan, Y. 2012. The Short of It: Investor Sentiment and Anomalies. *Journal of Financial Economics*. 104(2):288–302. DOI: 10.1016/j.jfineco.2011.12.001.

Statman, M. 2019. *Behavioural Finance: The Second Generation*. CFA Institute Research Foundation.

Steyn, D.H.W., Greyling, T., Rossouw, S. & Mwamba, J.M. 2020. Sentiment, Emotions and Stock Market Predictability in Developed and Emerging Markets. *GLO Discussion Paper, No. 502, Global Labor Organization (GLO), Essen*. 35.

Swamy, V. & Dharani, M. 2019. Google Search Intensity and the Investor Attention Effect: A Quantile Regression Approach. *Journal of Quantitative Economics*. 18:403–423. DOI: 10.1007/s40953-019-00185-9.

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. DOI: 10.1111/j.1540-6261.2007.01232.x.

Tetlock, P.C. 2011. All the News That's Fit to Reprint: Do Investors React to Stale Information? *Review of Financial Studies*. 24(5):1481–1512. DOI: 10.1093/rfs/hhq141.

Tetlock, P.C., Saar-Tsechansky, M. & Macskassy, S. 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*. 63(3):1437–1467. DOI: 10.1111/j.1540-6261.2008.01362.x.

Thaler, R. & Johnson, E.J. 1990. Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. *Management Science*. 36(6):643–660.

Thompson, D. 2011. *The World's First Twitter-Based Hedge Fund Is Finally Open for Business*. Available: https://www.theatlantic.com/business/archive/2011/05/the-worlds-first-twitter-based-hedge-fund-is-finally-open-for-business/239097/ [2020, September 05].

Twedt, B. & Rees, L. 2011. Reading Between the Lines: An Empirical Examination of Qualitative Attributes of Financial Analysts' Reports. *Journal of Accounting and Public Policy*. 31. DOI: 10.1016/j.jaccpubpol.2011.10.010.

Wysocki, P.D. 1998. *Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards*. (SSRN Scholarly Paper ID 160170). Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.160170.

Xu, Y., Liu, Z., Zhao, J. & Su, C. 2017. Weibo Sentiments and Stock Return: A Time-frequency View. *PLOS ONE*. 12(7):e0180723. DOI: 10.1371/journal.pone.0180723.

You, W., Guo, Y. & Peng, C. 2017. Twitter's Daily Happiness Sentiment and the Predictability of Stock Returns. *Finance Research Letters*. 23:58–64. DOI: 10.1016/j.frl.2017.07.018.

APPENDICES

Appendix A: List of companies

Table A1: List of companies included in the study

1	ABSA Group Limited	41	Dischem Pharmacies
2	ArcelorMittal SA Limited	42	Distell Group Holdings
3	Advtech Ltd	43	DrD Gold Limited
4	Allied Electronics Corp	44	Discovery Limited
5	Aeci Limited	45	Datatec Limited
6	Alexander Forbes Group Limited	46	Emira Property Fund
7	Afrimat Limited	47	Eoh Holdings
8	African Oxygen Limited	48	EPP N.V
9	Anglo American Plc	49	Equities Properties Fund
10	Arrowhead Prop Ltd	50	Exxaro Resources Ltd
11	African Rainbow	51	Famous Brands Ltd
12	Adcock Ingram Holdings Ltd	52	Fortress Reit Ltd A
13	Anglo American Platinum Ltd	53	Fortress Reit Ltd B
14	Anglogold Ashanti Ltd	54	Firstrand Ltd
15	Anheuser-Busch Inbev SA	55	Gold Fields Ltd
16	Accelerate Prop Fund Ltd	56	Glencore Plc
17	Aspen Pharmacare Holdings Ltd	57	Grindrod Ltd
18	African Rainbow Minerals Limited	58	Growthpoint Prop Ltd
19	Astral Foods Limited	59	Hamorny Gm Co Ltd
20	Ascendis Health Limited	60	Hoskins Cons Inv Ltd
21	Assore Limited	61	Hudaco Industries Ltd
22	Attacq Limited	62	Hammerson Plc
23	Avi Ltd	63	Hospitality Prop Fund B
24	Brait Se	64	Hyprop Invest Ltd Conv
25	Barloworld Ltd	65	Investec Australia Prop
26	BHP Group Ltd	66	Impala Platinum Holdings
27	Bid Corporation Ltd	67	Investec Limited
28	Blue Label Telecoms Ltd	68	Investec Plc
29	Brimstone Investments Ltd	69	Investec Property Fund
30	British American Tobacco	70	Imperial Logistics
31	Bidvest Ltd	71	Italtile Ltd
32	Capital and Counties	72	Intu Properties Plc
33	Compagnie Fin Richemont	73	Invicta Holdings Ltd
34	City Lodge Hotels Limited	74	Jse Ltd
35	Clicks Group	75	Kap Industrial Holdings Ltd
36	Coronation Fund Managers	76	Kumba Iron Ore Limited
37	Curro Holdings Ltd	77	PSG Consult Ltd
38	Capitec Bank Holdings	78	Liberty Two Degrees Ltd
39	Cashbuild Limited	79	Long 4 Life Ltd
40	Cartrack Holdings	80	Liberty Holdings Ltd

Continued

81	Libstar Holdings Ltd	120	Remgro Ltd
82	Lewis Group Ltd	121	Resilient Reit Ltd
83	Life Healthcare Group Holdings Ltd	122	Rfg Holdings
84	Lighthouse capital Ltd	123	Reunert Ltd
85	Multichoice Group Ltd	124	RMB Holdings Ltd
86	Mediclinic Int clinic	125	Rand Merchant Investments Ltd
87	Mondi Plc	126	Reinet Investments S.C.A
88	Mpact Ltd	127	Rdi Reit Plc
89	Mr Price Group Ltd	128	SA corp Real Estate Ltd
90	Massmart Holdings Ltd	129	Sappi Ltd
91	Mas Real Estate	130	Standard Bank Group
92	Metair Investments Ltd	131	Stadio Holdings Ltd
93	Motus Holdings Ltd	132	Shoprite Holdings
94	Momentum Met Holdings	133	Sanlam Ltd
95	Mtn Group Holdings	134	Steinhoff International Holdings
96	Murray & Roberts Holdings	135	Santam Ltd
97	Nedbank Group Ltd	136	Sasol Ltd
98	Northam Platinum Ltd	137	Super Group Ltd
99	Nampak Ltd	138	The Spar Group Ltd
100	Naspers Ltd	139	Sirius Real Estate Ltd
101	Nepi Rockcastle Plc	140	Stor-Age Pro Reit Ltd
102	Netcare Ltd	141	Sibanye Stillwater Ltd
103	Oceana Group Ltd	142	Stenprop Ltd
104	Octodec Invest Ltd	143	Sun International
105	Omnia Holdings	144	Spur Corporation Ltd
106	Old Mutual Ltd	145	Tiger Brands Ltd
107	Pan African Resource	146	Transaction Capital Ltd
108	Pioneer Goods Group Ltd	147	The Foschini Group Ltd
109	Peregrine Holdings Ltd	148	Tsogo Sun Hotels Ltd
110	Pick N Pay Stores Ltd	149	Telkom SA Ltd
111	Ppc Ltd	150	Trencor Ltd
112	Pepkor Holdings Ltd	151	Truworths Ltd
113	Prosus N.V	152	Tsogo Sun Gaming Ltd
114	PSG Group Ltd	153	Textainer Group Holdings
115	Quilter Ltd	154	Vukile Property Fund Ltd
116	Royal Bafokeng Platinum	155	Vodacom Group Ltd
117	Raubex Group Ltd	156	Wilson Bayly Hlm-Ovc Ltd
118	Rcl Foods Ltd	157	Woolworths Holdings Ltd
119	Redefine Properties Ltd	158	Zeder Inv Ltd

Appendix B: Panel unit root tests

Table B1: Eviews output from unit root tests using daily data

PANEL A

Panel unit root test: Summary
Series: BULLISHNESS
Date: 09/06/20 Time: 11:59
Sample: 1/05/2015 12/31/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 20
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-0.64212	0.0000	158	180516
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-60.8683	0.0000	158	180516
ADF - Fisher Chi-square	4286.43	0.0000	158	180516
PP - Fisher Chi-square	7459.50	0.0000	158	180516

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL B

Panel unit root test: Summary
Series: RETURNS
Date: 09/06/20 Time: 16:38
Sample: 1/05/2015 12/31/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 6
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-531.086	0.0000	158	170516
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-452.163	0.0000	158	170516
ADF - Fisher Chi-square	18562.3	0.0000	158	170516
PP - Fisher Chi-square	14024.1	0.0000	158	170516

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL C

Panel unit root test: Summary
Series: AGREEMENT
Date: 09/06/20 Time: 20:42
Sample: 1/05/2015 12/31/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 8
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-4.78998	0.0000	158	121716
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-46.8539	0.0000	158	121716
ADF - Fisher Chi-square	3586.33	0.0000	158	121716
PP - Fisher Chi-square	7834.67	0.0000	158	121716

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL D

Panel unit root test: Summary
Series: MESSAGES
Date: 09/06/20 Time: 20:56
Sample: 1/05/2015 12/31/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 4
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-24.0633	0.0000	158	156492
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-64.2794	0.0000	158	156492
ADF - Fisher Chi-square	5175.54	0.0000	158	156492
PP - Fisher Chi-square	11427.4	0.0000	158	156492

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

Table B2: Eviews output from unit root tests using weekly data

PANEL A

Panel unit root test: Summary
Series: BULLISHNESS
Date: 09/09/27 Time: 11:58
Sample: 1/07/2015 12/25/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 20
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-4.32415	0.0000	125	32078
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-20.8024	0.0000	125	32078
ADF - Fisher Chi-square	1584.79	0.0000	125	32078
PP - Fisher Chi-square	1950.38	0.0000	125	32078

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL B

Panel unit root test: Summary
Series: MESSAGES
Date: 09/09/20 Time: 11:00
Sample: 1/07/2015 12/25/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 6
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-36.1971	0.0000	132	34001
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-56.9962	0.0000	132	34001
ADF - Fisher Chi-square	4443.82	0.0000	132	34001
PP - Fisher Chi-square	6391.30	0.0000	132	34188

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL C

Panel unit root test: Summary
Series: AGREEMENT
Date: 09/10/20 Time: 11:06
Sample: 1/07/2015 12/25/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 6
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-5.91274	0.0000	126	32425
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-29.5134	0.0000	126	32425
ADF - Fisher Chi-square	2297.93	0.0000	126	32425
PP - Fisher Chi-square	2574.64	0.0000	126	32508

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL D

Panel unit root test: Summary
Series: RETURNS
Date: 09/27/20 Time: 11:37
Sample: 1/07/2015 12/25/2019
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 8
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-200.564	0.0000	157	33839
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-189.144	0.0000	157	33839
ADF - Fisher Chi-square	16710.5	0.0000	157	33839
PP - Fisher Chi-square	17392.1	0.0000	157	34045

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

Table B3: Eviews output from unit root tests using monthly data

PANEL A

Panel unit root test: Summary
Series: BULLISHNESS
Date: 09/27/20 Time: 11:37
Sample:2015M03 2019M12
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 4
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-115.803	0.0000	126	4662
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-28.3135	0.0000	126	4662
ADF - Fisher Chi-square	513.070	0.0000	126	4662
PP - Fisher Chi-square	607.004	0.0000	126	4788

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL B

Panel unit root test: Summary
Series: RETURNS
Date: 09/27/20 Time: 00:06
Sample: 2015M03 2019M12
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 4
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-3.97622	0.0000	144	2903
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-4.20289	0.0000	144	2903
ADF - Fisher Chi-square	390.008	0.0000	144	2903
PP - Fisher Chi-square	760.627	0.0000	144	2903

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL C

Panel unit root test: Summary
Series: AGREEMENT
Date: 09/27/20 Time: 00:09
Sample: 2015M03 2019M12
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 9
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-12.8878	0.0000	118	4354
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-17.7793	0.0000	118	4354
ADF - Fisher Chi-square	920.821	0.0000	118	4354
PP - Fisher Chi-square	1052.23	0.0000	118	4354

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

PANEL D

Panel unit root test: Summary
Series: MESSAGES
Date: 09/27/20 Time: 00:10
Sample: 2015M03 2019M12
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 8
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-18.7388	0.0000	115	4321
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-26.0641	0.0000	115	4321
ADF - Fisher Chi-square	1277.03	0.0000	115	4321
PP - Fisher Chi-square	1408.00	0.0000	115	4321

** Probabilities for Fisher tests are computed using an asymptotic Chi
-square distribution. All other tests assume asymptotic normality.

Appendix C: Robustness tests

Table C1: Random effects estimation

	Dependent Variable		
	Return	*Volatility*	*Trading volume*
	(1)	*(2)*	*(2)*
Bullishness	0.004***	1.086***	-0.0002
	(0.001)	(0.046)	(0.026)
Agreement	-0.004***	-1.647***	0.211***
	(0.001)	(0.008)	(0.004)
Messages	0.0001	-0.208***	0.084***
	(0.0001)	(0.008)	(0.004)
Market Return	0.615***	0.450	-0.867***
	(0.006)	(0.467)	(0.262)
Constant	-0.0003***	26.606***	13.014***
	(0.0001)	(0.587)	(0.119)
Observations	180 724	180 887	180 881
R^2	0.062	0.011	0.004
Adjusted R^2	0.062	0.010	0.004
F Statistic	11 993.42***	1 997.723***	437.945***
Notes:	**p<0.1; **p<0.05; ***p<0.01*		

Table C2: Quantile Regression with firm-fixed effects

	Dependent variable:				
	Stock Return				
	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
Bullishness	0.004***	0.003***	0.002***	0.001*	0.001
	(0.001)	(0.001)	(0.0004)	(0.001)	(0.001)
Agreement	-0.009***	-0.004***	-0.001*	0.005***	0.018***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Messages	-0.001***	-0.0002***	0.0001***	0.0004***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Market return	0.6602***	0.592***	0.501	0.603***	0.664***
	(0.008)	(0.0061)	(0.004)	(0.006)	(0.010)
Constant	-0.017***	-0.009***	-0.0004***	0.009***	0.022***
	(0.001)	(0.001)	(0.00004)	(0.0001)	(0.0001)
Pseudo R^2	0.042	0.045	0.031	0.043	0.038
Observations	180 724	180 724	180 724	180 724	180 724
Notes:	**p<0.1; **p<0.05; ***p<0.01*				

Table C3: 1-day lag Granger non-causality tests

PANEL A

Pairwise Dumitrescu Hurlin Panel Causality Tests
Date: 09/06/20 Time: 23:20
Sample: 1/05/2015 12/31/2019
Lags: 1

Null Hypothesis:	W-Stat.	Zbar-Stat.	Prob.
BULLISHNESS does not homogeneously cause RETURNS	1.3739	0.59600	0.0653
RETURNS does not homogeneously cause BULLISHNESS	3.0226	3.28960	9.E-37

PANEL B

Pairwise Dumitrescu Hurlin Panel Causality Tests
Date: 09/12/20 Time: 18:44
Sample: 1/05/2015 12/31/2019
Lags: 1

Null Hypothesis:	W-Stat.	Zbar-Stat.	Prob.
MESSAGES does not homogeneously cause RETURNS	1.37396	0.40442	0.6859
RETURNS does not homogeneously cause MESSAGES	3.02263	3.28960	0.0010

PANEL C

Pairwise Dumitrescu Hurlin Panel Causality Tests
Date: 09/12/20 Time: 18:13
Sample: 1/05/2015 12/31/2019
Lags: 1

Null Hypothesis:	W-Stat.	Zbar-Stat.	Prob.
RETURNS does not homogeneously cause AGREEMENT	3.25551	3.69715	0.0002
AGREEMENT does not homogeneously cause RETURNS	1.56739	0.74294	0.4575

Table C4: 2-day lag Granger non-causality test

PANEL A

Pairwise Dumitrescu Hurlin Panel Causality Tests
Date: 09/06/20 Time: 16:22
Sample: 1/05/2015 12/31/2019
Lags: 2

Null Hypothesis:	W-Stat.	Zbar-Stat.	Prob.
BULLISHNESS does not homogeneously cause RETURNS	2.89903	0.59600	0.1481
RETURNS does not homogeneously cause BULLISHNESS	3.87569	1.68320	1.E-21

PANEL B

Pairwise Dumitrescu Hurlin Panel Causality Tests
Date: 09/06/20 Time: 18:59
Sample: 1/05/2015 12/31/2019
Lags: 2

Null Hypothesis:	W-Stat.	Zbar-Stat.	Prob.
MESSAGES does not homogeneously cause RETURNS	2.87905	0.56700	0.5512
RETURNS does not homogeneously cause MESSAGES	3.87564	1.88620	0.0023

PANEL C

Pairwise Dumitrescu Hurlin Panel Causality Tests
Date: 09/06/20 Time: 18:59
Sample: 1/05/2015 12/31/2019
Lags: 2

Null Hypothesis:	W-Stat.	Zbar-Stat.	Prob.
RETURNS does not homogeneously cause AGREEMENT	4.82435	2.73923	0.0062
AGREEMENT does not homogeneously cause RETURNS	2.48819	0.13865	0.8897

Table C5: 1-week lag Granger non-causality tests

PANEL A

Pairwise Granger Causality Tests
Date: 09/25/20 Time: 23:56
Sample: 1/07/2015 12/25/2019
Lags: 1

Null Hypothesis:	Obs	F-Statistic	Prob.
BULLISHNESS does not Granger Cause RETURNS	34045	0.99394	0.3188
RETURNS does not Granger Cause BULLISHNESS		1.04690	0.3062

PANEL B

Pairwise Granger Causality Tests
Date: 09/25/20 Time: 23:56
Sample: 1/07/2015 12/25/2019
Lags: 1

Null Hypothesis:	Obs	F-Statistic	Prob.
MESSAGES does not Granger Cause RETURNS	34045	0.01522	0.9018
RETURNS does not Granger Cause MESSAGES		0.52626	0.4682

PANEL C

Pairwise Granger Causality Tests
Date: 09/25/20 Time: 00:26
Sample: 1/07/2015 12/25/2019
Lags: 1

Null Hypothesis:	Obs	F-Statistic	Prob.
AGREEMENT does not Granger Cause RETURNS	34045	0.97085	0.3245
RETURNS does not Granger Cause AGREEMENT		0.00235	0.9614

Table C6: 2-week lag Granger non-causality tests

PANEL A

Pairwise Granger Causality Tests
Date: 09/25/20 Time: 00:26
Sample: 1/07/2015 12/25/2019
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
BULLISHNESS does not Granger Cause RETURNS	32874	1.96459	0.1445
RETURNS does not Granger Cause BULLISHNESS		2.63324	0.0719

PANEL B

Pairwise Granger Causality Tests
Date: 09/25/20 Time: 00:32
Sample: 1/07/2015 12/25/2019
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
MESSAGES does not Granger Cause RETURNS	32874	0.42453	0.6541
RETURNS does not Granger Cause MESSAGES		0.82600	0.4378

PANEL C

Pairwise Granger Causality Tests
Date: 09/25/20 Time: 00:57
Sample: 1/07/2015 12/25/2019
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
AGREEMENT does not Granger Cause RETURNS	32874	2.98096	0.0508
RETURNS does not Granger Cause AGREEMENT		0.52005	0.5945

Table C7: 1-month lag Granger non-causality tests

PANEL A

Pairwise Granger Causality Tests
Date: 09/26/20 Time: 20:42
Sample: 2015M03 2019M12
Lags: 1

Null Hypothesis:	Obs	F-Statistic	Prob.
RETURNS does not Granger Cause BULLISHNESS	3238	2.5E-06	0.9987
BULLISHNESS does not Granger Cause RETURNS		0.25854	0.6112

PANEL B

Pairwise Granger Causality Tests
Date: 09/26/20 Time: 20:42
Sample: 2015M03 2019M12
Lags: 1

Null Hypothesis:	Obs	F-Statistic	Prob.
MESSAGES does not Granger Cause RETURNS	3238	1.12450	0.2890
RETURNS does not Granger Cause MESSAGES		0.81981	0.3653

PANEL C

Pairwise Granger Causality Tests
Date: 09/26/20 Time: 20:47
Sample: 2015M03 2019M12
Lags: 1

Null Hypothesis:	Obs	F-Statistic	Prob.
RETURNS does not Granger Cause AGREEMENT	3238	0.00014	0.9907
AGREEMENT does not Granger Cause RETURNS		0.21646	0.6418

Table C8: 2-month lag Granger non-causality tests

PANEL A

Pairwise Granger Causality Tests
Date: 09/26/20 Time: 20:47
Sample: 2015M03 2019M12
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
BULLISHNESS does not Granger Cause RETURNS	2915	0.21604	0.8057
RETURNS does not Granger Cause BULLISHNESS		0.02114	0.9791

PANEL B

Pairwise Granger Causality Tests
Date: 09/26/20 Time: 20:47
Sample: 2015M03 2019M12
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
MESSAGES does not Granger Cause RETURNS	2915	3.07479	0.0563
RETURNS does not Granger Cause MESSAGES		0.27405	0.7603

PANEL C

Pairwise Granger Causality Tests
Date: 09/26/20 Time: 20:48
Sample: 2015M03 2019M12
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
RETURNS does not Granger Cause AGREEMENT	2915	0.60647	0.5453
AGREEMENT does not Granger Cause RETURNS		0.04940	0.9518

Table C9: Quantile estimators using $\tau = \epsilon[0.05, 0.2, 0.4, 0.6, 0.8, 0.95]$

	Dependent variable:					
	Stock Return					
	$\tau = 0.05$	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.6$	$\tau = 0.8$	$\tau = 0.95$
Bullishness	0.00568***	0.003731***	0.002548***	0.001558***	0.00118*	0.001648
	(0.00122)	(0.00064)	(0.000485)	(0.000470)	(0.0007)	(0.001125)
Agreement	-0.03861***	-0.00969***	-0.003793***	0.000484	0.006544***	0.02868***
	(0.005876)	(0.001303)	(0.000854)	(0.000786)	(0.00145)	(0.003508)
Messages	-0.00121***	-0.000245***	-2.46E-05	0.000272***	0.000523***	0.0014***
	(0.000268)	(8.52E-05)	(5.15E-05)	(5.00E-05)	(7.77E-05)	(0.000225)
Market return	0.68038***	0.624798***	0.5274***	0.523454***	0.625838***	0.65953***
	(0.014273)	(0.006071)	(0.005505)	(0.005567)	(0.006093)	(0.015203)
Constant	-0.03127***	-0.013063***	-0.00362***	0.002941***	0.01248***	0.03119***
	(0.000175)	(7.01E-05)	(4.65E-05)	(4.58E-05)	(7.01E-05)	(0.00017)
Pseudo R^2	0.037218	0.045458	0.039686	0.038259	0.043378	0.029836
Observations	180 724	180 724	180 724	180 724	180 724	180 724

Note*:* **p<0.1; **p<0.05; ***p<0.01*

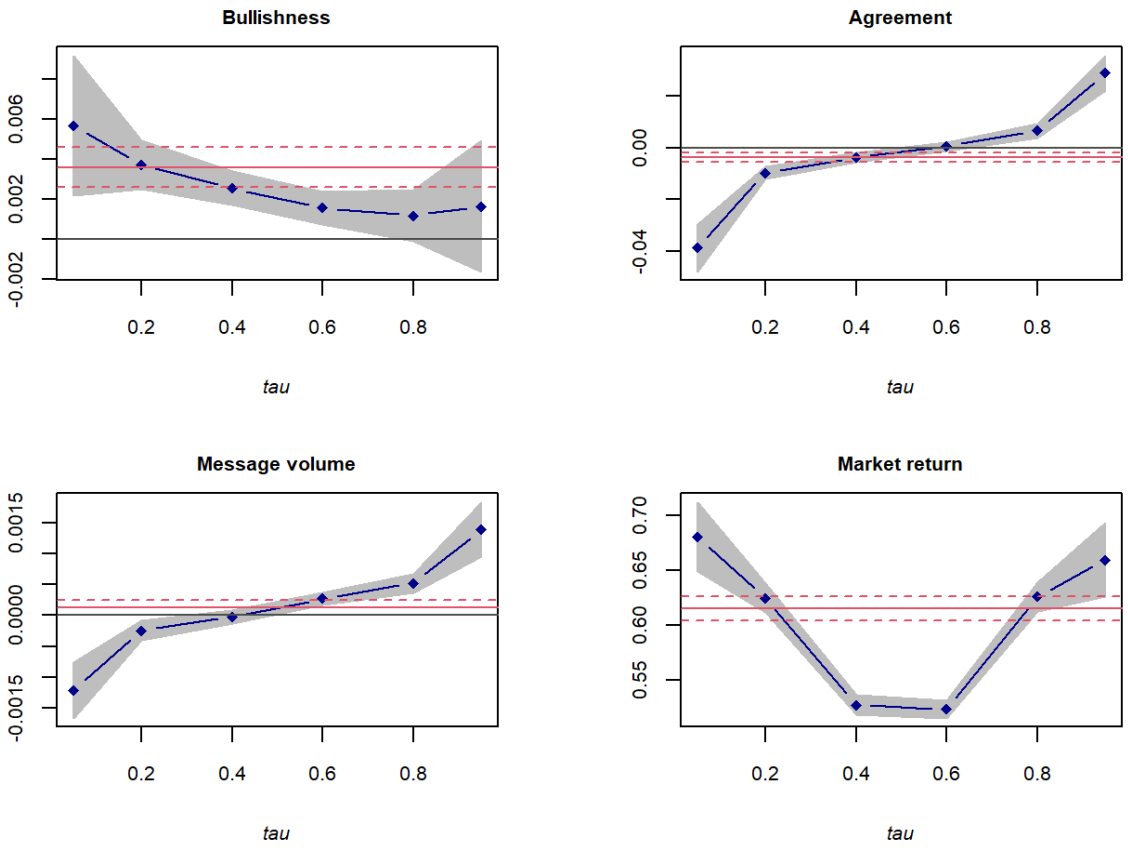


Figure C1: Visualisation of the quantile estimators using τ=ϵ[0.05,0.2,0.4,0.6,0.8,0.95]

Notes: *The red continuous line shows the Ordinary Least Squares (OLS) estimate while the two dotted lines show the 95% confidence interval of the OLS estimates and the brown continuous line shows the point where the vertical axis is equal to zero. The vertical axis shows the coefficients for the bullishness score at the various quantiles while the horizontal axis indicates the quantile distributions of the returns* ($\tau\epsilon[0.1, 0.25, 0.5, 0.75, 0.9]$). *The values of the estimated* $\beta(\tau)$ *parameters are connected by the blue line while the grey shaded area indicates the 95% confidence intervals of the* $\beta(\tau)$ *estimated parameters.*

Appendix D: R-code for Quantile regression

Libraries

```
rm(list = ls())        #Clearing the global environment
library(stargazer)     #For exporting regression output to pdf
library(quantreg)      #For the quantile regression estimates
library(readxl)        #Importing from excel
library(tidyverse)
```

Importing the data from excel

```
dataset <- read_csv("combined_zeroxl.csv")
```

Quantile regressions

rq_0.1 stands for quantile regression at quantile 0.1 etc

```
rq_0.1 <-
  rq(Stock_return ~ Av_tweet + Agreement + Msg_vol + Market_return,
     tau = 0.1,
     dataset)

rq_0.25 <-
  rq(Stock_return ~ Av_tweet + Agreement + Msg_vol + Market_return,
     tau = 0.25,
     dataset)

rq_0.5 <-
  rq(Stock_return ~ Av_tweet + Agreement + Msg_vol + Market_return,
     tau = 0.5,
     dataset)

rq_0.75 <-
  rq(Stock_return ~ Av_tweet + Agreement + Msg_vol + Market_return,
     tau = 0.75,
     dataset)

rq_0.9 <-
  rq(Stock_return ~ Av_tweet + Agreement + Msg_vol + Market_return,
     tau = 0.9,
     dataset)
```

Exporting the regression results as latex

```
stargazer(
  rq_0.1,
  rq_0.25,
  rq_0.5,
  rq_0.75,
  rq_0.9,
  covariate.labels = c("Bullishness", "Agreement",
                       "Messages Volume", "Market Return"),
  dep.var.labels = "Stock Return",
  column.labels = c("$\\tau=0.1$",
                    "$\\tau=0.25$",
                    "$\\tau=0.5$",
                    "$\\tau=0.75$",
                    "$\\tau=0.9$"),
  type = "latex",
  model.numbers = FALSE
  )
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Mon, Sep 14, 2020 - 10:59:52 PM