Investor Beliefs and their Impact on Financial Markets

Dissertation to obtain the doctoral degree of Economic Sciences (Dr. oec.)

Faculty of Business, Economics and Social Sciences

University of Hohenheim

Institute of Financial Management

submitted by

Carolin Hartmann

from *Stuttgart*

2021

Day of the oral examination: 11.05.2022

Dean of the Faculty of Business, Economics, and Social Sciences at the University of Hohenheim: Prof. Dr. Jörg Schiller

First Reviewer: Prof. Dr. Hans-Peter Burghof

Second Reviewer: Prof. Dr. Dirk Hachmeister

Contents

List of Abbreviations

API	Application Programming Interface
AR	Autoregressive
BNP	Banque Nationale de Paris
CBOE	Chicago Board Options Exchange
DALAHO	Data Laboratory University of Hohenheim
DAX	Deutscher Aktienindex (German Share Index)
DJIA	Dow Jones Industrial Average Index
DSSW	Model of Delong, Shleifer, Summers and Waldmann
DZ Bank	Deutsche Zentral-Genossenschaftsbank
e.g.	For Example
ECB	European Central Bank
EONIA	European Over Night Index Average
et al.	And Ohers
EUREX	European Exchange
EURIBOR	Euro Interbank Offered Rate
FED	Federal Reserve
FSE	Frankfurt Stock Exchange
GSV	Google Search Volume

HML	High Minus Low
IRF	Impulse Response Functions
Lcaps	Large-Caps
LL	Log Likelihood
MAE	Mean Absolut Error
Max.	Maximum
Mcaps	Mid-Caps
Min.	Minimum
MOM	Monthly Momentum Factor
N	Number of Observations
NASDAQ	National Association of Securities Dealers Automated Quotations
NYSE	New York Stock Exchange
OLS	Ordinary Least Square
RBS	Royal Bank of Scotland
RDB	Research Database
RF	Risk-Free Rate
RM	Market Return
RMRF	Market Excess Return
RMSE	Root Mean Squared Error
RV	Realized Volatility

S&P	Standard & Poor's
Scaps	Small-Caps
S.D.	Standard Deviation
SE	Standard Error
SEC	Securities and Exchange Commission
SIRCA	Securities Industry Research Centre of Asia-Pacific
SMB	Small Minus Big
SSE	Stuttgart Stock Exchange
T3	3-Month Treasury Bill
TV	Twitter Volume
US	United States
USD	United States Dollar
Var	Variance
VAR	Vector Autoregression
VDAX	DAX- Volatility Index
VIX	CBOE Volatility Index
VSSE	SSE Volatility Index
VXD	DJIA Volatility Index
XETRA	Exchange Electronic Trading

List of Symbols

α	Constant
β_S	Sentiment Coefficient
β_X	Coefficient of the set of control variables
β_E	Coefficient of EONIA
β_T	Coefficient of the DAX turnover
$\beta_{1,\dots n}$	Regression Coefficient
γ	Risk aversion
Δ	Percentage change
$\Delta GSV_{i,t}$	Percentage change in GSV of company i at time t
$\Delta TV_{i,t}$	Percentage change in TV of company i at time t
$\Delta Turnover_{i,t}$	Percentage change in Turnover of company i at time t
$\Delta Volatility_{i,t}$	Percentage change in Volatility of company i at time t
$\Delta Squared\_Returns_{i,t}$	Percentage change in Squared Returns of company i at time t
ε_t	Error term
θ	Constant
$\hat{\theta}^c$	Corrected estimator of the constant
θ	Constant
$\hat{\theta}^c$	Corrected estimator of the constant
λ_t	Time fixed effect

μ	Share of noise traders
$1-\mu$	Share of sophisticated investors
μ_i	Cross-sectional fixed effect
ρ_t	Misperception
ρ	Sentiment coefficient
$\hat{\rho}$	Estimator of the sentiment coefficient
$\hat{\rho}^c$	Corrected estimator of the sentiment coefficient
σ^2	Variance
ϕ	Coefficient of the unobservable error term v_t
Gr_t	GSV residual
$G2r_t$	GSV residual controlling for TV
i	Sophisticated investor
n	Noise trader
N_i	Total number of shares outstanding of company i at time t
p_t	Price of the risky asset
%	Percent
$R_{i,t}$	Portfolio return with $i = Smallcaps, Largecaps, Bigcaps$ at time t
$r_{t,j}^2$	Squared returns at day t for the interval j
r	Interest rate
$(1+r)$	Discount factor with interest rate r

s	Investment in a safe asset
t	Time period with $t = 1, \dots T$
Tr_t	TV residual
$T2r_t$	TV residual controlling for GSV
u	Investment in a risky asset
u_t	Error term at time t
v_t	Unobservable error term
v_t^c	Corrected proxy variable of the unobservable error term
$v_{i,t}$	Error term for company i at time t
X_t	Set of control variables
$X_{\mathrm{i,t}}$	Volume of shares traded of company i with at time t
$x_{i,t}$	Independent variable $1xk$ vector of company i at time t
$y_{i,t}$	Dependent variable for company i at time t

List of Tables

List of Figures

1 Introduction

In financial theory the behaviour of individuals plays an important role. One of the best-known approaches is to assume that individuals behave rationally, as the Efficient Market Hypothesis suggests. Rational behaviour is predictable and therefore easy to model. Reality shows that the assumption of rationality is often not sufficient to reflect the behaviour of individuals on financial markets. Individual investors have their own beliefs regarding market figures and the development of financial markets. It is difficult to measure their beliefs and therefore their effect on financial markets such as return and volatility.

One option is to approximate investor beliefs with publicly available information in newspapers and on websites. New data sources open up possibilities to complement these known measures. There are for example granular trade data, Google Search Volume and Twitter tweets that are available on a daily and intra-daily level. They all have one thing in common, a closer observation of individual behaviour. It is possible to track what and when individuals trade. Google shares insights on what individuals search for. Twitter provides information on what people tweet and retweet.

The idea of this work is to use these new data sources to approximate investor beliefs. It investigates whether the approximation improves the measurement of return and volatility in existing model frameworks. Further, it assesses the impact of investor beliefs on the forecasting of return and volatility. The aim is to deepen the understanding of individual investors, their behaviour and the impact of their beliefs on financial markets.

This thesis unites three articles, presented in the following chapters.The second chapter “Battle of investors on the German market – A Sentiment Analysis” was written with Professor Patrick Roger from the University of Strasbourg. The third chapter “Google and Twitter Data: Two Perspectives on the Behaviour of Noise Traders” is a collaborative work with Professor Dr. Hans-Peter Burghof of the University of Hohenheim and Professor Dr. Marc Mehlhorn from the University of Applied Sciences in Cologne. The fourth chapter “The impact of Twitter and Google on Volatility - A Time-Series Analysis” is a single authored paper.

In chapter 2 of this thesis, we compare the trading behaviour of individual investors and institutional investors to see if the behaviour of individual investors predicts market movements. Therefore, we build a sentiment index using the difference of implied volatilities between warrants on the German Share Index DAX, traded on the retail-oriented Stuttgart Stock Exchange (SSE), and option contracts on the European Exchange EUREX, mainly traded by professionals in Frankfurt. Over a four-year period, we show that our index is significant in predicting the daily returns of a size-based long-short portfolio of stocks. We reinforce this result by controlling for Fama-French factors. To verify the result we conduct several robustness checks, such as using implied volatilities from call and put warrants traded on the SSE or implied volatilities based on buy and sell orders. All the results show that our sentiment indicator is significant in predicting returns. Our results confirm that retail and professional investors behave differently in a systematic way which is reflected in persistent mispricing.

In chapter 3, we use GSV and TV to make the beliefs of individual investors measurable. We analyse their impact on trading activity and volatility in a panel data set-up over a period of 3.5 years. We measure and compare the impact of percentage changes in GSV and TV on financial markets. We find that online investor sentiment measured by GSV and TV have an impact on financial markets and predictive power. First, an increase in GSV and TV has a positive impact on trading activity on the same and on the next day which is in line with Easley et al. (1996) indicating that new market entrants leads to higher trading activity. Second, an increase in TV leads to an increase in volatility, following the DSSW model by De Long et al. (Long, Shleifer, Summers, & Waldmann, 1990) that the share of noise traders increase. Changes in GSV have no significant impact on volatility and thus on the share of noise traders in the market. Our overall results are robust to a variety of tests. Including market liquidity as a control variable to the regression set-up does not change the impact of GSV and TV on trading activity. To control for endogeneity issues due to the integration of the lagged variables in the regression on trading activity and volatility, we apply the Arellano-Bond estimator which does not affect our results.

In chapter 4, I examine the effects of investor attention and investor sentiment on realized volatility (RV) of the Dow Jones Industrial Average (DJIA) over a period of 2.5 years. For this empirical study, I use a time series set-up with data from GSV and TV. The basis forms a standard Vector autoregression

(VAR) model with an exogenous variable to take macroeconomic and financial factors into account. In order to clearly identify the effect of GSV and TV on RV, I use the residuals of GSV and TV controlling for the macroeconomic and financial factors. In line with existing literature, it turns out that both have an impact on the RV of the DJIA. Although, the effect of TV on RV is more important as it is significant at the 1 % level while GSV is only significant at the 5 % level. For the in-sample forecasting, the linear prediction model with GSV and TV residuals outperforms a standard AR (1) process. Out-of-sample the AR (1) process outperforms the standard model with GSV and TV residuals. Their influence remains significant but small. Clustering for high and low volatility groups, the analysis shows that the effect of GSV and TV on RV changes. Especially in times of high and low RV, GSV and TV seem to contain new information, as they improve the model fit compared to a standard AR (1) process. However, the results are not persistent in- and out-of-sample. Using a rolling window approach for the forecasting does not change the results. Same holds for the change of the forecasting period.

In chapter 5, I conclude on the findings of the previous chapters and give an outlook. Particularly with regard to investor sentiment and attention measures, there are recent developments that combine a variety of new internet platform data including GSV and TV, as well as new model approaches such as machine learning and neural networks.

Battle of investors on the German market – A Sentiment Analysis[1]

Carolin Hartmann[2] and Patrick Roger[3]

[1] The authors thank Eurofidai (www.eurofidai.org) for the availability of time series data on Fama-French Factors. We also thank the DALAHO of the University of Hohenheim for the availability of the intraday and daily market data on stock prices from Sirca and Datastream. We are particularly grateful for the trading dataset of the Stuttgart stock exchange. Our thanks also go to the INEF exchange program of the DAAD between the University of Strasbourg and the University of Hohenheim which made this research project possible. We also appreciated the comments of Brian A. Thompson and Steven Li, and participants at the Behavioural Finance Working Group Meeting, London 2018, the 11th Academy of Behavioral Finance and Economics meeting, Chicago 2018, and the 36th French Finance Association meeting, Québec 2019. This paper circulated previously under the title "Is Stuttgart more sentiment prone than Frankfurt? The case of retail oriented structured products in Germany"

[2] University of Hohenheim, Institute of Financial Management, Mailing address: Schwerzstraße 38, 70599 Stuttgart, GERMANY, Tel: +49 (0) 7 11 45 92 29 00, Email: carolin_hartmann@uni-hohenheim.de

[3] LaRGE Research Center, EM Strasbourg Business School, university of Strasbourg, Mailing address: 61 avenue de la Forêt Noire, 67085 Strasbourg Cedex, FRANCE, Tel: +33 (0) 3 68 85 21 56, Email: proger@unistra.fr

2 Battle of investors on the German market – A Sentiment Analysis

2.1 Introduction

Retail and professional investors have different beliefs and do behave differently in many situations, especially when trading risky assets[4]. In the 1980s, retail investors were identified as noise traders because "they trade on noise as if it were information" (Black, 1986). In efficient markets, noise traders cannot influence prices. Since the first papers by Terrance Odean (1998, 1999), however, the literature suggests that this is not the case. The main stylized facts of this literature are 1) retail investors hold underdiversified portfolios, 2) they narrowly frame their decisions, and 3) they often trade in the same direction, therefore influencing market prices.

Underdiversification of retail investors' portfolios was highlighted by Lease, Lewellen, and Schlarbaum (1974) and Blume and Friend (1975), and confirmed by more recent studies (Broihanne, Merli, & Roger, 2016; Goetzmann & Kumar, 2008; Kumar, 2007; Mitton & Vorkink, 2007). Theoretical papers justify underdiversification by the desire of investors to hold positively skewed portfolios (Barberis & Huang, 2008; Brunnermeier, Gollier, & Parker, 2007; Brunnermeier & Parker, 2005).

The narrowly framed decisions of retail investors lead them to evaluate stocks in isolation (Barberis, Huang, & Thaler, 2006). Contrary to standard expected utility theory and Markowitz's portfolio choice theory, narrow framing means that retail investors do not consider their portfolio as a whole. Trading a given security is motivated by optimism/pessimism about the future return on this specific financial security.

The third stylized fact, i.e. correlated trading, can move prices and drive future returns, as illustrated in Dorn, Huberman, and Sengmueller (2008). With a sample of 37,000 clients of a German broker, they show that investors' trades are correlated in a systematic way. These authors find that correlated limit

[4] The anecdotal evidence of the bitcoin futures contracts illustrates this well-known phenomenon. In December 2017, bitcoin futures were launched on the Chicago Board Options Exchange (Cboe, 2017; CFTC, 2017). The analysis of trades during the first month showed that retail investors were much more bullish on bitcoins than professional investors. In short, large traders took short positions and retail investors held long positions (Ahmed, 2018; Cboe, 2018; Osipovich, 2018). Since then, the bitcoin price sharply decreased before partially recovering. After a peak of 19,843 USD on 16 Dec 2017 the price of the bitcoin decreased to less than 4,000 USD at the beginning of 2019, before bouncing back over 10,000 USD in July 2019.

orders have some predictive power of subsequent market returns. Kumar and Lee (2006) also find that stocks heavily held by retail investors comove more together, than they comove with other stocks.

The three individual characteristics described above can generate persistent mispricing. This makes the construction of a sentiment index that can have a significant predictive power of future returns, especially relevant. In fact, optimism/pessimism of investors is often translated in terms of investor sentiment, which is defined by Baker and Wurgler (Baker & Wurgler, 2007) as "a belief about future cash flows and investment risks that is not justified by the facts at hand".

In the finance literature, professional investors are often presented as rational people taking decisions based on fundamentals, when retail investors are more prone to sentiment trading[5] (the so-called noise traders according to Fischer Black (1986)). Put in different words, retail investors perceive a distorted distribution of future prices and returns, not in line with the facts at hand (Baker & Wurgler, 2007; De Long, Shleifer, Summers, & Waldmann, 1990).

In this framework, our aim is twofold. First, we build a measure of investor sentiment based on the activity of retail investors on structured products traded on the Stuttgart Stock Exchange (henceforth SSE), namely warrants whose underlying is the DAX30 performance index. Structured products on the SSE are tailored for a clientele of retail investors. Therefore, the implied volatility extracted from prices of these products is a "retail volatility" that can be compared to the implied volatility based on the EUREX option contracts, mainly traded by professional investors in Frankfurt. This Frankfurt implied volatility is provided by EUREX under the name VDAX-NEW in the recent years, and VDAX before. VDAX and VDAX-NEW are the German equivalents of the VIX index on the Chicago Board Options Exchange (CBOE).

Our sentiment index, defined as the difference between the implied volatilities[6] on the two markets, has a predictive power of size-based portfolio returns. We apply the standard methodology (Baker & Wurgler, 2006) to demonstrate this predictive power on a daily basis.

[5] Most of the literature argues that sentiment-induced mispricing is caused by retail investors (Barberis and Xiong (2012), Da et al. (2015), Lee et al. (1991), Baker and Wurgler (2006, 2007)). It should be noted, however, that this view has been recently challenged by DeVault et al. (2019); these authors show on a number of sentiment metrics that these metrics mainly capture demand shocks of institutional investors.
[6] Denoted VSSE for the warrants traded on the Stuttgart Stock Exchange and VDAXNEW for the options traded on the EUREX in Frankfurt.

Our analysis shows that retail investors over- or under-react (compared to professional investors) to events on the German market. This result is in the spirit of Poteshman's paper (2001). Poteshman finds that "investors underreact to information over short horizons and overreact to information over long horizons".

The contributions of our work come from the specificity of the SSE which is unambiguously retail-oriented. Our index measures as directly as possible a difference in beliefs between retail and professional investors. The warrants traded on the SSE are issued by large financial institutions that ensure the liquidity of the market. This microstructure effect could weaken our results. Therefore, we perform a number of robustness checks to eliminate alternative explanations. We control for a potential issuer effect and for a number of other confounding effects, for example the link between liquidity and maturity of contracts[7], the initiator of the trade (buyer/seller), or the impact of some policy decisions by the European Central Bank (ECB).

This chapter is organized as follows. Section 2 develops the sentiment indices based on the implied volatility differences between the two markets. Section 3 presents to data and descriptive statistics. Section 4 develops the main results and section 5 the robustness tests.

2.2 A sentiment index based on implied volatility

In this chapter, we focus on warrants and option contracts on the main German blue chip index "DAX performance index". The DAX index, usually named DAX 30, is an equity portfolio of the 30 largest German companies (accounting for 80 % of the German market capitalization). All DAX stocks are listed in the German prime standard and continuously traded on XETRA. The DAX30 is a free float market cap weighted index. The DAX performance index assumes that all dividends of the DAX30 firms are fully reinvested in the index (Deutsche Börse AG, 2007). In 2018, 28.9 million index option contracts were traded on the DAX at the EUREX.

[7] The SSE defines itself as an exchange for retail investor trading (Boerse Stuttgart, 2018a) and the trading of derivatives contracts takes place at EUWAX. Contrary to EUREX, the SSE has a hybrid market model with electronic and human trading (but no algorithmic trading and a minimum order size of one unit) (Boerse Stuttgart, 2018b). The latter (EUREX) publishes an implied volatility index based on option prices traded on EUREX.

Deutsche Börse AG calculates two option-based implied volatility indices called VDAX[8] and VDAX-NEW. The VDAX is based on the Black-Scholes-Model and is published once a day. The VDAX-NEW is continuously calculated on the options traded at the EUREX. This index is the underlying of highly liquid futures and option contracts.

2.2.1 VDAX and VDAX-NEW

The VDAX is calculated at 5:45 pm everyday as an average of the implied volatilities of the at-the-money call and put option prices traded during the day. The VDAX value is the annualized implied volatility, linearly interpolated to a standardized maturity of 45 days. (Deutsche Börse AG, 2007)[9]

As mentioned before, the aim of the VDAX-NEW is to provide an underlying index for futures and option contracts on volatility. It is calculated every minute from 9:15 am to 5:30 pm. Without entering the details of the methodology (described in the document provided by Deutsche Börse AG (2007)), the VDAX-NEW is a synthetic index built from the best bid and ask option prices traded at the EUREX. Finally, the VDAX-NEW is the implied volatility of a synthetic option with a standardized 30-days time to maturity. Though calculation details differ, definitions and methodologies used for the VDAX and the VDAX-NEW are comparable. As a consequence, we will calculate the implied volatility of SSE contracts in the same way.

2.2.2 Implied volatility of SSE warrants

The SSE does not provide an implied volatility index. As mentioned above, we calculate a volatility index (denoted by VSSE) from DAX30 warrants traded at the SSE, using the methodology leading to the VDAX. The only difference lies in the input data (warrant prices instead of option prices). In a first step, we calculate the implied volatility of all warrants using the Black-Scholes model. To be consistent with the hypotheses of the Black-Scholes model, we restrict the sample to European-style warrants on the DAX performance index. We therefore get a per trade implied volatility.

[8] The VDAX is available until the 29 July 2016.
[9] The new manual on the DAX is available since December 2018. Link to the most recent source https://www.dax-indices.com/resources. Our calculations are based on the manual of 2007 which is valid for the period we consider in this chapter.

In a second step, we calculate the daily implied volatility index (VSSE) on the DAX30 using our per trade implied volatility. We exclude warrants with a time to maturity smaller or equal to ten days, to avoid random fluctuations of implied volatility in the last trading days. The daily VSSE index is the median of the per trade implied volatilities on a given day.

2.2.3 The sentiment indicator

The sentiment indicator on day t, denoted $Sentiment_t$, is the difference between $VDAX_t$ and $VSSE_t$.

$$Sentiment_t = VSSE_t - VDAX_t \tag{2.1}$$

The sentiment indicator increases if the VSSE increases and/or the VDAX decreases. It decreases if the VSSE decreases and/or the VDAX increases. In other words, sentiment reinforces when retail investors are ready to pay warrants more than the price professionals are ready to pay the corresponding options.

2.3 Data and Descriptive statistics

2.3.1 Dataset

Our sample covers the period April 2009 - July 2013. We use five information sources: 1) the research database (RDB) of the SSE 2) the intraday data from SIRCA[10] by Thomson Reuters, 3) daily values of the German Fama-French factors provided by Eurofidai, 4) daily values of the German Fama-French factors provided by Brückner et al. (2015), and 5) interest rates data from the German federal bank.

The RDB of the SSE contains two main files; master and trading data. The former identifies the tradable products at the SSE; the latter contains all the data about orders and trades. The combination of these two datasets allows us to identify the traded warrants on the DAX entering our calculation of the VSSE.

Overall there were 24.6 million executed orders on 5.8 million products at the SSE in the sample period. Focusing on European-style warrants on the DAX, 35,775 products remain in the sample. Among them,

[10] Securities Industry Research Centre of Asia-Pacific (SIRCA) provides access to vast and comprehensive online repositories of global news and financial markets data. The data is provided by the DALAHO of the Univesity of Hohenheim.

11,492 were traded in the sample period.[11] For these traded warrants we find 135,126 executed orders. In the sample database, each executed order stands for one observation.

The DAX index

The DAX values are taken from the database SIRCA, on a one-minute basis. The data comes from the Frankfurt Stock Exchange (FSE) XETRA. We merge each order at the SSE with the DAX value of the corresponding minute.[12] However, trading hours differ on the two markets. SSE works from 8:00 am to 10:00 pm, and XETRA works from 9:00 am to 5:30 pm at the FSE, (the last price is published at 5:45 pm). In our sample, SSE trades after 5:45 pm are merged with the closing DAX value and SSE trades executed before 9:00 am are merged with the 9:00 am DAX value.

VDAX and VDAX-NEW

Similar to the DAX values, the intraday values of the VDAX-NEW are available on SIRCA.[13] The availability of the VDAX-NEW data is limited to a time span between 9:15 am and 5:30 pm (it was the same range for the DAX). Thus, we merge the values in the morning with the first VDAX-NEW value and the trades after 5:30pm with the last VDAX-NEW value of the day.

The merging process of the VDAX with the database is easier, as only one daily value is defined. Each observation is merged to the VDAX of the trading day of each warrant.

The Fama-French factors

Eurofidai (www.eurofidai.org) provides the size portfolios (terciles) and the Fama-French factors for the German market. In section 5, we perform a robustness test with a second set of German factors calculated by Brückner et al. (2015).

2.3.2 Sample

Our final dataset contains 135,126 warrant trades for which we know the order number, the time-stamp of each trade and the trading price. The average time to maturity of the traded warrants is 174 days. The

[11] Though all products are registered in the master data file, they were not necessarily traded during the observation period. This can be due to multiple exercise prices and maturities.
[12] If there is no DAX value in the minute of the trade due to a lack of data, we take the more recent available value.
[13] The initial source is the FSE.

median value of a trade is 3,045€. 45 % (55 %) of trades are purchases (sales). Trades are quite balanced between calls and puts; 46 % (54 %) of the traded warrants are calls (puts). Over the sample period, the 3-month Euribor stayed within the range [0.19 %; 1.62 %]. Table 2.1 provides the monthly trading volumes over the period of study.

Table 2.1 Yearly and monthly trading volumes

Number of trades (per year and month)													
	Jan	Feb	March	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
2009				2,203	1,744	1,453	1,980	1,542	1,275	1,571	1,893	1,105	14,766
2010	1,326	1,633	2,789	3,171	5,262	2,339	2,134	2,382	2,341	2,653	3,210	1,230	30,470
2011	1,383	1,678	2,886	1,368	2,338	1,637	1,993	8,331	5,770	4,944	6,312	3,449	42,089
2012	3,726	3,432	3,011	2,584	3,886	3,489	4,089	3,749	2,914	2,819	2,632	1,480	37,811
2013	1,739	1,913	1,439	1,615	1,438	1,090	756						9,990
Total	8,174	8,656	10,125	10,941	14,668	10,008	10,952	16,004	12,300	11,987	14,047	7,264	135,126

Number of contracts traded between April 2009 and July 2013.

There are 1,104 trading days over the period, and 142 daily trades on average. The minimum (maximum) trading volume is 13 (1006) contracts. The number of trades is evenly distributed between the different days of the week.

Panel A of Table 2.2 provides the distribution of trades within a typical day. Not surprisingly, the distribution is U-shaped with less trades at lunch time. 9.98 % of trades take place after the closing of XETRA at 5:30 pm. Panel B shows that warrants are issued by 11 different issuers but the four main issuers, namely BNP Paribas, Vontobel, Commerzbank and DZ Bank, represent 86 % of the trades.

Table 2.2 Distribution of trades during the day (Panel A) and issuers of warrants (Panel B)

Panel A: Distribution of trades over the day					
Time	Nr. of trades	Percent	Time	Nr. of trades	Percent
0:00 - 09:00	2793	2 %	3:00 - 4:00	13516	10 %
9:00 - 10:00	26294	19 %	4:00 - 5:00	16156	12 %
10:00 - 11:00	14644	11 %	5:00 - 6:00	10641	8 %
11:00 - 12:00	11652	9 %	6:00 - 7:00	4918	4 %
12.00 - 1:00	9109	7 %	7:00 – 8:00	4638	3 %
1:00 - 2:00	8920	7 %	8:00 - 9:00	0	0 %
2:00 - 3:00	11845	9 %	9:00 - 10:00	0	0 %
Panel B: Issuers of warrants in the sample					
Issuer	Frequency	Percent	Issuer	Frequency	Percent
BNP	55,405	41.00 %	UBS	5,204	3.85 %
Vontobel	20,634	15.27 %	Lang & Schwarz	2,997	2.22 %
Commerzbank	20,616	15.26 %	RBS	1,514	1.12 %
DZ Bank	19,007	14.07 %	Société Générale	23	0.02 %
Goldman Sachs	9,721	7.19 %	Deutsche Bank	5	0.00 %

The moneyness of a call (put) trade is equal to the ratio of the DAX value (strike price) and the strike price (DAX value). The moneyness of the trades in our sample ranges between 0.305 and 2.467 with an average of 0.925. 78 % (22 %) of the trades are out-of-the money (in-the-money) trades. The number of call (73,278) and put (61,848) trades is almost balanced.

2.3.3 Summary statistics

We first calculate the VSSE. Overall there are 40 trading months that is 1104 trading days. Table 2.3 reports time series statistics on three different volatility indices and the sentiment indicator. On a daily level, the VSSE, the VDAX and the VDAX-NEW show slightly different mean values. The VDAX exhibits the lowest mean (22.101 %), followed by the VSSE (23.885 %) and the VDAX-NEW (24.1885 %). The implied volatility we observe at the SSE is comparable to what we find at the FSE. The distribution of the indices over the period is close to normal, the VSSE being the closest with a skewness coefficient of 0.66 and a kurtosis almost equal to the theoretical value of 3. The VDAX and the VDAX-NEW are more peaked and positively skewed. It could be a signal of over-reaction on the FSE, consistent with the findings of DeVault et al. (2019). The sentiment indicator (*Sentiment*), has a positive mean (1.755 %). The negative minimum value (-11.336 %) indicates that the VSSE is some-

times sharply smaller than the VDAX. The standard deviation (2.519 %) shows that the sentiment indicator is not stable over time. This is not a surprising result, even if the period under scrutiny was almost always bullish.

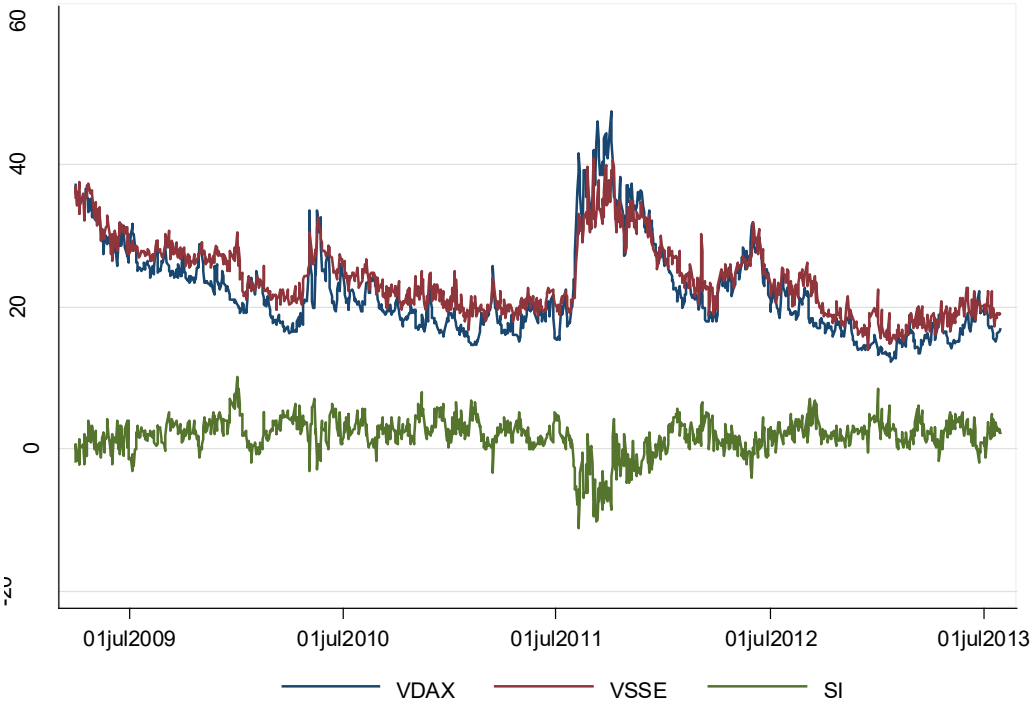
Table 2.3 Summary statistics on implied volatility indices

	VSSE	VDAX	VDAX-NEW	Sentiment
Mean	23.8851	22.1011	24.1751	1.7551
Median	22.9043	20.46	22.2985	1.98483
S.D.	5.00062	6.34261	7.17002	2.51863
Min.	13.7444	12.29	13.23	-11.3362
Max.	40.7595	47.3	51.0067	10.1253
Skewness	0.66274	1.12829	1.1721	-1.16925
Kurtosis	3.01011	4.00286	4.13917	6.34573

VSSE denotes the SSE implied volatility index, VDAX and VDAX-NEW are the implied volatility measures of the Frankfurt stock exchange. The sentiment indicator is the difference between VSSE and VDAX. The VDAX was calculated until 29 July 2016 on a daily basis, the VDAX-NEW is continuously calculated during the trading day. Figures are given in percent.

Figure 2.1 shows the time series of implied volatilities and sentiment. The sentiment indicator is positive most of the time with a VSSE lying above the VDAX. Nevertheless, there are sub-periods where the VDAX lies above the VSSE, for example during the mid-2011 crisis during which the VDAX increased more than the VSSE. It illustrates that some market events (a severe market drop occurred at that time) lead to different reactions from retail vs. professional investors.

Figure 2.1 Sentiment indicator and implied volatilities
(April 2009 - July 2013)



2.4 Empirical study

2.4.1 Correlation analysis

One of the important findings of Baker and Wurgler (2006) is that large firms are less affected by sentiment than small firms. A low-sentiment period is usually followed by higher subsequent returns on small stocks (compared to large stocks) and a high-sentiment period does not necessarily generate an impact on returns. Baker and Wurgler (2007) conclude that stocks of smaller companies are more sensitive to sentiment than "bond-like" stocks which tend to be less affected by investor sentiment. The authors summarize the consequence of their findings in the "seesaw sentiment". In high-sentiment periods, larger stocks can be undervalued and small stocks overvalued. The reason is that arbitrage on large stocks is more difficult than arbitrage on small stocks. Based on these assumptions, D'Hondt and Roger (2017) show that small caps are overvalued (undervalued) compared to large caps in high-sentiment (low-sentiment) periods. According to the authors, the quality of a sentiment indicator depends on its ability to forecast future returns. Moreover, the correlation of a good sentiment indicator should be higher with the future returns on small stocks than with the future returns on large stocks.

Table 2.4 reports two types of correlations based on daily data. The Table contains two rows and eight columns. The first four columns relate to the four Fama-French-Carhart factors, namely the market excess return (RMRF), size (SMB), book-to-market (HML) and momentum (MOM). The last four columns relate to the terciles of size-based portfolios, namely Large-caps (Lcaps), Mid-caps (Mcaps), and Small-caps (Scaps), and to the difference between the Small caps portfolio and the Large caps portfolio (last column of Table 2.4). The first row of the Table gives the correlations between the sentiment indicator and the eight variables just defined above. The second row gives the correlations between the lagged (1 day lag) sentiment indicator and the same eight variables. The Fama-French-Carhart factors and the size-based portfolio returns come from Eurofidai[14].

Table 2.4 Correlation analysis

	RMRF	MOM	HML	SMB	Scaps	Mcaps	Lcaps	Small-Big
$Sentiment_t$	0.287***	-0.056	0.097***	-0.221***	0.212***	0.270***	0.282***	-0.214***
$Sentiment_{t-1}$	-0.035	0.009	-0.027	0.113***	0.167***	0.173***	-0.038	0.108***

Correlation between sentiment indicator, Fama-French factors and portfolios. The market sentiment indicator is the daily difference between the VSSE and the VDAX. The four Fama-French-Carhart factors are: the market return (RMRF), the size factor (SMB), the value factor (HML) and the momentum factor (MOM). They are calculated as the corresponding factors on the US market. The three portfolios Scaps, Mcaps and Lcaps represent the portfolio returns based on size terciles (Scaps for small-caps, Mcaps for mid-caps and Lcaps for large-caps). The factors and portfolios are provided by Eurofidai for the German market. The column "Small-Big" stands for the difference between the small cap portfolio returns and the large cap portfolio return (Scaps – Lcaps). Variables are calculated on a daily basis from April 2009 to July 2013. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

The correlation of the sentiment indicator $Sentiment_t$ and the four Fama-French-Carhart factors is highly significant at the 1 % level for all factors, except momentum. Consistent with Roger (2014), we find a highly significant positive correlation between the $Sentiment_t$ and the market return (0.2867). If the market return increases, the sentiment indicator increases as well. In our case, this means that the difference between the VSSE and the VDAX gets larger indicating that the investors at the SSE and the FSE react differently to variation of market returns. The book-to-market factor (HML) is positively correlated with the sentiment indicator (0.0973) as well and the size factor (SMB) is negatively correlated with the sentiment indicator (-0.2205). All three size portfolios (small-caps, mid-caps and large-

[14] For the German market, Eurofidai provides value-weighted portfolios which are split up in three size portfolios, small-cap (size 1), mid-cap (size 2) and large-cap (size 3), according to their market capitalization.

caps) are positively correlated with the sentiment indicator. The return of the size-based long-short portfolio (Small-Big) is negatively correlated with the sentiment indicator (-0.2139). One possible interpretation of the negative signs is given by the disposition effect. If prices fall retail investors tend to keep losing stocks or buy new stocks. When prices increase investors tend to sell their stocks too early which leads to a decrease of the sentiment indicator.

For the lagged sentiment indicator $Sentiment_{t-1}$, only the correlation with the size factor (SMB) remains highly significant at the 1 % level and positive (0.113). In line with this observation the correlation between the lagged sentiment indicator and the Small-Big portfolio returns is also positive and highly significant at the 1 % level (0.108). For the different size-based portfolios, the correlation is positive and significant at the 1 % level for the small caps (0.167) and the mid-caps (0.173). As expected, the correlation between the sentiment indicator and the large caps is not significant. It confirms that smaller stocks are more affected by changes in the lagged sentiment indicator. These findings on the relevance of the different size-based portfolios are in line with D'Hondt and Roger (2017).

2.4.2 Multivariate analysis

In this section, we follow the standard methodology of Baker and Wurgler (2006) to test the predictive power of our sentiment index. The dependent variable is the daily return of a long-short portfolio based on size. In our case, this is the difference between the return on the portfolio of small caps (first tercile), denoted $R_{Smallcaps,t}$, and the return on the portfolio of large caps (third tercile), denoted $R_{Largecaps,t}$.

In a first step, we perform a univariate regression to test whether the lagged sentiment indicator is significantly linked to the portfolio return[15].

$$R_{Smallcaps,t} - R_{Largecaps,t} = \alpha + \beta_S Sentiment_{t-1} + \varepsilon_t \quad (2.2)$$

In a second step, we control for the Fama-French-Carhart factors.

$$R_{Smallcaps,t} - R_{Bigcaps,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_t + \varepsilon_t \quad (2.3)$$

[15] Though this simple regression looks redundant after presenting the correlation analysis, we keep it because it will become useful in some robustness checks, especially when we will control for autocorrelation of the sentiment index (section 5.5).

The vector X stands for the set of control variables, namely the market excess return (RMRF), the book-to market factor (HML), and the momentum factor (MOM). As in Baker and Wurgler (2006), we exclude the SMB factor in the regression analysis. In fact, the correlation between the size factor (SMB) and the dependent variable is (not surprisingly) equal to 0.96.

Table 2.5 presents the results of the basic regression set-up. Column (1) stands for the univariate equation (2) without control variables. Column (2) stands for equation (3), including the Fama-French-Carhart factors from Eurofidai for the German market as control variables. Column (3) also stands for the results of equation (3), including the Fama-French-Carhart factors from Brückner et al. (2015) for the German market. The sentiment coefficient β_S is positive and significant for the three equations. Our sentiment indicator is therefore statistically significant for the prediction of future returns.

Table 2.5 Basic regression with French data

VARIABLES	(1) R_{small}-R_{big}	(2) R_{small}-R_{big}	(3) R_{small}-R_{big}
$Sentiment_{t-1}$(x10^4)	4.75**	3.35***	2.69***
	(3.595)	(5.053)	(4.011)
$RMRF_t$		-0.872***	-0.825***
		(-50.16)	(-54.67)
HML_t		-0.0154	0.0290
		(-0.552)	(1.176)
MOM_t		0.0503*	0.0741***
		(1.942)	(3.723)
Constant (x10^4)	-8.99**	-1.27	-0.14
	(-2.216)	(-0.618)	(-0.0728)
N	1,094	1,094	1,094
R^2	0.012	0.855	0.858

Regressing the return of a size-based long-short portfolio on lagged sentiment and Fama-French factors with Eurofidai data. Column (1) reports the results of the simple regression without control variables: $R_{Smallcaps,t} - R_{Largecaps,t} = \alpha + \beta_S Sentiment_{t-1} + \varepsilon_t$. Column (2) stands for the equation with control variables: $R_{Small,t} - R_{Bigc,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_t + \varepsilon_t$. The control variables are the Fama-French-Carhart factors from Eurofidai for the German market. The third column represents the same equation with Fama-French-Carhart factors from Brückner et al. (2015) for the German market (daily ALL). The SMB factor is not included, as the long-short portfolio is based on this criterion. The sentiment indicator is calculated for the whole sample on a daily basis. Without the sentiment indicator the R^2 is 0.8477. We control for heteroscedasticity by using heteroscedasticity-consistent standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

The results in column (1) show a positive and highly significant impact of the sentiment indicator $Sentiment_{t-1}$ on the long-short portfolio. The coefficient β_S is positive (4.75 x 10^{-4}) and significant at the 5 % level. This means that a period of high (low) sentiment is followed by periods of high (low) returns on the long-short portfolio. The return on the portfolio of small firms increases and these firms tend to be

overvalued, compared to the larger firms. The positive coefficient points out that an increase in the difference between the two investor groups at the SSE and the FSE in $t-1$ has a positive impact on the return of the long-short portfolio $R_{Smallcaps,t} - R_{Largecaps,t}$. The investors behave differently on the two stock exchanges; this difference generates mispricing.

In Baker and Wurgler (Baker & Wurgler, 2007), Roger (2014) and D'Hondt and Roger (2017) the sentiment coefficient was negative. One reason for the sign difference with our coefficient is the definition of our indicator, based on a difference of two implied volatilities. In particular, a higher VSSE means that investors in Stuttgart are ready to "overpay" warrants, compared to option prices in Frankfurt. As warrants include calls and puts, we can only link the sentiment indicator to the optimism of investors to the future return of their *own* strategy. In other words, retail investors who buy call (put) warrants are confident, maybe overconfident, that the market index will increase (drop) in the near future.

Column (2) reports the results of equation (3), including the Fama-French-Carhart factors as control variables. The coefficient of the sentiment indicator $Sentiment_{t-1}$ is positive and significant at the 1 % level (3.35×10^{-4}). As expected, the coefficient of the market excess return (RMRF) is negative and highly significant (-0.872). This result is standard because the portfolio of small caps is much less correlated to the market excess return than the portfolio of large caps. The book-to-market coefficient (HML) is positive but insignificant and the momentum coefficient (MOM) is positive but only marginally significant at the 10 % level. The high R^2 (0.855) comes mainly from the market excess return and marginally from the sentiment index. It is then remarkable to keep the lagged sentiment significant in the controlled version of the regression.

The results in column (3) are obtained with the Fama-French-Carhart factors of Brückner et al. (2015) . To calculate the Fama-French-Carhart factors, the authors adapt the usual method to the peculiarity of the German market. In Table 2.5, we use their *daily all* dataset[16]. The results are similar to those in column (2). The coefficient of the sentiment indicator is slightly lower but is still significant at the 1 %

[16] Here is the definition given by the authors for the two market segments they consider over the period we study in this paper.
1) TOP – all stocks in the Regulated Market that were formerly listed in the Official Market. We do not include firms listed in the middle segment of the FSE (Regulated Market) and in its New Market.
2) ALL – all German stocks listed in the Regulated Market of the FSE.

level. The significant coefficients RMRF and MOM are also smaller but keep the same sign, compared to column (2). Finally, the R^2 of the regression is virtually unchanged at 0.858. In the next section, we consider, as a robustness check, the Fama-French-Carhart factors of Brückner et al. (2015) calculated from the top segment of the German market.

2.5 Robustness tests

We now perform a number of robustness tests to check whether our previous results are contingent on a set of methodological choices. First, as mentioned in footnote 16, we use another set of Fama-French-Carhart factors from Brückner et al. (2015). This alternate set of factors is based on a reduced set of stocks, namely the top segment of the German market. Second, we aim at testing a pure predictive regression by introducing the lagged values of the Fama-French-Carhart factors. Third, we separate buy and sell orders in the regression analysis. Fourth, we use different definitions of our sentiment indicators to test the sensitivity of our results to slight modifications of our indicator. Fifth, we take into account the possible autocorrelation of the sentiment indicator, as in Roger (2014) and Roger and D'Hondt (2017). Sixth, we use the EONIA as a proxy variable controlling for the ECB policy and the DAX turnover to control the impact of market movements on our results.

2.5.1 Fama-French-Carhart factors

As mentioned before, Brückner et al. (2015) provide a second dataset of Fama-French-Carhart factors for the regulated German market which they recommend for studies on stocks listed in the top segment. As the DAX is the underlying index for the VDAX and the VSSE and covers the top 30 listed companies in Germany, we use this set of factors to confirm our results. The SMB and HML factors are constructed in line with Fama and French (1993). The momentum factor (MOM) of Carhart (1997) is calculated based on Fama & French (2012). The results are reported in Table 2.6.

Overall, the R^2 increases by 0.86 % but results are virtually unchanged. The major difference lies in the HML factor which becomes significant at the 5 % level (0.0401)[17]. The coefficient of the sentiment indicator slightly decreases but remains significant at the 1 % level.

Table 2.6 Basic regression with German data

VARIABLES	(1) R_{small}-R_{big}
$Sentiment_{t-1}$(x10^4)	2.53***
	(3.850)
$RMRF\_TOP_t$	-0.804***
	(-56.57)
HML_TOP_t	0.0401**
	(2.110)
MOM_TOP_t	0.071***
	(4.098)
Constant(x10^4)	-0.013
	(-0.007)
N	1,094
R^2	0.861

Regressing the return of a size-based long-short portfolio on lagged sentiment and Fama-French factors from Brückner et al., 2015 leads to the coefficient of sentiment in the simple regression: $R_{Smallcaps,t} - R_{Bigcaps,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_t + \varepsilon_t$ controlling for the market return (RMRF), the Fama-French factor (HML) and the Carhart momentum factor (MOM). We use the dataset from Brückner et al. (2015) on the top segment for the German market as control variables. The top segment contains all stocks of firms that are listed on the Regulated Market. The authors do not include firms listed in the middle segment of the FSE (regulated market) and in its new market. The SMB factor is not included as the long-short portfolio is based on this criterion. The sentiment indicator is calculated for the whole sample on a daily basis. Without the sentiment indicator the R^2 of the equation is 0.8564. We control for heteroscedasticity by using heteroscedasticity-consistent standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

2.5.2 Lagged control variables

In section 4, we followed Baker and Wurgler (Baker & Wurgler, 2007) for the introduction of control variables in the regression. The long-short portfolio returns were regressed on the lagged sentiment indicator and the contemporaneous control variables (the excess market return, RMRF, and the Fama-French-Carhart factors, HML and MOM). Roger (2014) also tested a pure predictive regression in which the control variables are also lagged (X_{t-1} instead of X_t). As a robustness check we also test equation (4) below. The results are presented in table 2.7.

$$R_{Small,t} - R_{Big,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_{t-1} + \varepsilon_t \quad (2.4)$$

[17] The HML factor for the top (whole) German market is on average at 0.038% (0.0463%). Not surprisingly, the difference between value and growth stocks is less pronounced in the top segment than on the whole German market.

The sentiment indicator remains statistically significant but the control variables are not significantly different from zero. It means that the control variables have no explanatory power in equation (2.4).

Table 2.7 Basic regression with lagged Fama-French factors

VARIABLES	(1) R_{small}-R_{big}
$Sentiment_{t-1}$(x10^4)	5.04**
	(2.349)
$RMRF_{t-1}$	-0.0266
	(-0.733)
HML_{t-1}	0.0144
	(0.193)
MOM_{t-1}	-0.0179
	(-0.316)
Constant(x10^4)	-9.27
	(-1.526)
N	1,094
R^2	0.012

Regressing the return of a size-based long-short portfolio on lagged sentiment and lagged Fama-French factors give the coefficient of sentiment in the simple regression: $R_{Smallcaps,t} - R_{Bigcaps,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_{t-1} + \varepsilon_t$ controlling for the Fama-French factors and the Carhart momentum factor using the data from Eurofidai. The sentiment indicator and the control variables are lagged variables in the regression. The sentiment indicator is calculated for the whole sample on a daily basis. The SMB factor is not included as the long-short portfolio is based on this criterion. Without the sentiment indicator the R^2 of the equation is 0.0003. We control for heteroscedasticity by using heteroscedasticity-consistent standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

2.5.3 Sentiment indicator built separately from buy and sell trades

The VSSE is calculated from a set of buy and sell orders. The issuers of the warrants are also market makers, thus proposing their own bid/ask prices. Of course, ask prices are above bid prices, therefore generating different implicit volatilities. We calculate separately the implied volatility using buy and sell orders at the SSE. Over the period, the two average implicit volatilities are very close to each other. The mean value for the Buy (Sell)-VSSE is 23.92 % (23.91 %). As a consequence, we calculate separate sentiment indicators for buy and sell orders. The mean values of the sentiment indicator based on each set of orders (buy/sell) are almost equal at 1.79 %.

$$Sentiment\_buy_t = VSSE\_buy_t - VDAX_t \quad (2.5)$$

$$Sentiment\_sell_t = VSSE\_sell_t - VDAX_t \quad (2.6)$$

We introduce the sentiment variables defined in equations (2.5) and (2.6) in the regressions of equations (2.2) and (2.3). The results appear in table 2.8. The lagged sentiment is still positive and significant in

all cases. Without control variables, the regression coefficient of lagged sentiment is lower for buy orders than for sell orders. This ranking is reversed in the controlled version of the regression but controlling for Fama-French-Carhart factors (RMRF, HML, MOM) does not change the results. In columns (3) and (6) we use the risk factors from Brückner et al. (2015). Coefficients for the lagged sentiment are slightly less significant but the results are qualitatively unchanged.

Table 2.8 Basic regression with lagged sentiment calculated from buy/sell orders

	BUY			SELL		
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	R_{small}-R_{big}	R_{small}-R_{big}	R_{small}-R_{big}	R_{small}-R_{big}	R_{small}-R_{big}	R_{small}-R_{big}
$Sentiment_{t-1}$(x10^4)	3.13*	3.26***	2.33***	5.13***	1.98***	1.82***
	(2.611)	(5.573)	(3.971)	(3.894)	(3.044)	(2.767)
$RMRF_t$		-0.876***	-0.827***		-0.871***	-0.825***
		(-50.96)	(-54.98)		(-48.40)	(-53.15)
HML_t		-0.0118	0.0293		-0.0186	0.0277
		(-0.426)	(1.189)		(-0.661)	(1.120)
MOM_t		0.0512**	0.0736***		0.0496*	0.0733***
		(1.983)	(3.704)		(1.874)	(3.630)
Constant (x10^4)	-6.25	-1.23	0.415	-9.80**	1.08	1.32
	(-1.575)	(-0.629)	(0.224)	(-2.407)	(0.522)	(0.678)
N	1,094	1,094	1,094	1,094	1,094	1,094
R^2	0.006	0.856	0.858	0.014	0.851	0.856

Regressing the return of a size-based long-short portfolio on lagged sentiment calculated from buy/sell orders give the coefficient of sentiment for buy and sell orders. The lagged sentiment indicator in column (1)-(3) is the difference between the VSSE for buy orders and the VDAX. In column (4)-(6) the sentiment indicator is the difference between VSSE for sell orders and the VDAX. In column (1) and (4) the univariate regression conducted: $R_{Smallcaps,t} - R_{Bigcaps,t} = \alpha + \beta_S Sentiment_{t-1} + \varepsilon_t$. In the columns (2), (3), (5) and (6) the underlying regression is $R_{Smallcaps,t} - R_{Bigcaps,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_{t-1} + \varepsilon_t$. In column (2) and (5) the control variables are the the Fama-French factor (HML), the market factor (RMRF) and the Carhart momentum factor (MOM) from Eurofidai for the German market. In columns (3) and (6) the data comes from Brückner et al. (2015). The sentiment indicator is calculated for the whole sample on a daily basis. The SMB factor is not included, as the long-short portfolio is based on this criterion. Without the sentiment indicator the R^2 of the equation is 0.8477 in column 2 and 5, 0.8533 in column 3 and 6. We control for heteroscedasticity by using heteroscedasticity-consistent standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

2.5.4 Alternative definitions of the sentiment indicator

So far we used the difference between the implied volatility index VSSE and the VDAX as the sentiment indicator. We introduce now two variations of this standard measure as a robustness check. The alternative measures are denoted Sentiment2 and Sentiment3. Sentiment2 is the difference between the VSSE and the VDAX-NEW. As the VDAX is no longer available on the market we want to know whether our results are consistent with the new implied volatility measure on the market. To obtain a daily value of the VDAX-NEW we use the median per day. Sentiment3 is the difference between the winsorized VSSE (1 % of values on each side are deleted) and the VDAX. Our daily measure of VSSE

is the median value of the set of values of VDANEW during the given day. In short, Sentiment2 and Sentiment3 are defined in equations (2.7) and (2.8).

$$Sentiment2_t = VSSE_t - VDAXNEW_t \tag{2.7}$$

$$Sentiment3_t = VSSE\_98_t - VDAX_t \tag{2.8}$$

The results are reported in table 2.9. The equations show that our results do not change, even if we use alternative sentiment indicators. Sentiment2 is highly significant (at the 1 % level with a coefficient equal to 3.16 x 10^{-4}). It is slightly smaller than the coefficient of the univariate case for the standard sentiment indicator and the R^2 decreases from 0.012 to 0.008. Including the Fama-French-Carhart factors (HML and MOM) and the market return (RMRF) as control variables leads to a decrease of the lagged sentiment indicator but the significance level remains at 1 %. For the lagged Sentiment3, the coefficient β_S is only 0.5 x 10^{-4} smaller than for the standard sentiment indicator. In the controlled case, the coefficient is again comparable to the results from table 2.5.

For both equations with control variables in column (2) and (4), the market return (RMRF) coefficient is negative and highly significant. The book-to market factor (HML) is not significant and the momentum factor (MOM) is significant but only at the 10 % level.

Table 2.9 Basic regression with an alternative definitions of sentiment

	Sentiment2		Sentiment 3	
	(1)	(2)	(3)	(4)
VARIABLES	R_{small}-R_{big}	R_{small}-R_{big}	R_{small}-R_{big}	R_{small}-R_{big}
$Sentiment_{t-1}$(x10^4)	3.16***	1.91***	4.70***	3.07***
	(3.013)	(3.514)	(3.733)	(4.849)
$RMRF_t$		-0.870***		-0.872***
		(-49.39)		(-49.87)
HML_t		-0.0224		-0.0157
		(-0.803)		(-0.565)
MOM_t		0.0497*		0.0498*
		(1.889)		(1.920)
Constant (x10^4)	0.067	5.01***	-8.59**	-0.586
	(0.0202)	(3.721)	(-2.177)	(-0.294)
N	1,103	1,103	1,094	1,094
R^2	0.008	0.851	0.013	0.854

Regressing the return of a size-based long-short portfolio on lagged sentiment using alternative definitions of sentiment give the coefficient of sentiment. The lagged sentiment indicator in column (1) and (2) is the Sentiment2, the daily difference between the VSSE and the VDAX-NEW. For the daily VDAX we take the median of the intraday values of the VDAX-NEW. In columns 3 and 4 the lagged Sentiment3 measures the difference between VSSE98 and VDAX. The

VSSE98 is calculated as the daily median of the VSSE intraday values, leaving out the highest and lowest 1 % of observations and keeping the remaining 98 %. In column (1) and (3) the simple regressions equations are: $R_{Smallcaps,t} - R_{Bigcaps,t} = \alpha + \beta_S Sentiment_{t-1} + \varepsilon_t$. In columns (2) and (4) the underlying regression is $R_{Smallcaps,t} - R_{Bigcaps,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_t + \varepsilon_t$ in which the control variables are the the Fama-French factor (HML), the market factor (RMRF) and the Carhart momentum factor (MOM) from Eurofidai for the German market. The sentiment indicator is calculated for the whole sample on a daily basis. The SMB factor is not included, as the long-short portfolio is based on this criterion. We control for heteroscedasticity by using heteroscedasticity-consistent standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

2.5.5 Autocorrelation of the sentiment indicator

Earlier studies from D'Hondt and Roger (2017) and Roger (2014) show that the sentiment regression coefficient β_S can be biased if the sentiment indicator follows an autoregressive process. As a result, the significance of the sentiment indicator could be overstated. We apply the bias reduction technique of Amihud and Hurvich (2004) to manage this problem. The authors implement a bias-corrected estimator based on the approach by Stambaugh (1999). They integrate a measure for the unobservable error term v_t of the AR(1) in the standard OLS regression. Equation (2.2) becomes:

$$R_{Small,t} - R_{Big,t} = \alpha + \beta_S Sentiment_{t-1} + \phi v_t + \varepsilon_t \quad (2.9)$$

v_t cannot be directly measured. Amihud and Hurvich (2004) propose a four-step procedure to find a proxy for v_t. First, they regress $Sentiment_t$ on $Sentiment_{t-1}$, assuming the sentiment indicator follows an AR (1) process.

$$Sentiment_t = \theta + \rho Sentiment_{t-1} + v_t \quad (2.10)$$

Second, based on the OLS estimator $\hat{\rho}$ of the above equation, the corrected estimator $\hat{\rho}^c$ is constructed (with $n = 1084$ in our case):

$$\hat{\rho}^c = \hat{\rho} + \frac{(1+3\hat{\rho})}{n} + \frac{3(1+\hat{\rho})}{n^2} \quad (2.11)$$

Third, the proxy variable for v_t is constructed, using the corrected estimator $\hat{\rho}^c$ and the estimator of the constant $\hat{\theta}^c$ from equation (2.11):

$$v_t^c = Sentiment_t - (\hat{\theta}^c + \hat{\rho}^c Sentiment_{t-1}) \quad (2.12)$$

The proxy v_t^c is introduced in the regressions (2) and (3). In the univariate case the equation becomes

$$R_{Small,t} - R_{Big,t} = \alpha + \beta_S Sentiment_{t-1} + \phi v_t^c + \varepsilon_t \quad (2.13)$$

In the controlled case the regression equation becomes

$$R_{Small,t} - R_{Big,t} = c + \beta_S Sentiment_{t-1} + \beta_X X_t + \phi v_t^c + \varepsilon_t \quad (2.14)$$

Finally, we adjust the t-stats with the corrected standard error of β_S. The correction is necessary because the estimator of β_S is downward biased. This approach is conducted for all regressions. The results are presented in table 2.10.

$$\widehat{SE}^c(\beta_S) = \sqrt{\hat{\phi}^2 Var(\hat{\rho}^c) + \widehat{SE}^2(\beta_S)} \quad (2.15)$$

For all three equations the coefficients of the lagged sentiment indicator stay positive and highly significant at the 1 % level. In the univariate regression in column (1) the coefficient increases to 5.12 x 10^{-4} and the R^2 is 0.219. Controlling for Fama-French-Carhart factors in columns (2) and (3), does not change the results which stay close to the ones obtained in Table 2.5.

Table 2.10 Basic regression controlling for the autocorrelation of sentiment

VARIABLES	(1) R_{small}-R_{big}	(2) R_{small}-R_{big}	(3) R_{small}-R_{big}
$Sentiment_{t-1}$(x10^4)	5.12***	3.31***	2.64***
	(4.365)	(4.986)	(3.903)
v_t^c (x10^4)	-29.6***	-1.59	-0.259
	(-12.22)	(-1.583)	(-0.244)
$RMRF_t$		-0.861***	-0.00822***
		(-44.20)	(-46.31)
HML_t		-0.0102	0.000274
		(-0.363)	(1.114)
MOM_t		0.0495*	0.000713***
		(1.892)	(3.572)
Constant (x10^4)	-10.1**	-1.10	0.138e
	(-1.997)	(-0.529)	(0.0694)
N	1,086	1,086	1,086
R^2	0.219	0.854	0.856

Regressing the return of a size-based long-short portfolio on lagged sentiment and Fama-French factors, controlling for the autocorrelation of sentiment give the coefficient of sentiment. The first column presents the univariate regression $R_{Small,t} - R_{Big,t} = \alpha + \beta_S Sentiment_{t-1} + \phi v_t^c + \varepsilon_t$. The second and third column provide the same regression including the Fama-Fench-Carhart factors as control variables $R_{Smallcaps,t} - R_{Bigcaps,t} = c + \beta_S Sentiment_{t-1} + \beta_X X_t + \phi v_t^c + \varepsilon_t$. In the second column we include the Fama-French-Carhart factors from Eurofidai on Germany, in the third column we include the factors of Brückner et al. (2015) for the German market.

2.5.6 ECB policy impacts and DAX turnover

The ECB policy and the interest rates changed during the time period of our study. These changes could partly explain our dependent variable. For example, a change in interest rates impacts expectations, and possibly risk aversion, which in turn impact the composition of portfolios. To take these interest rate changes into account, we use the European Over Night Index Average (EONIA) as a control variable. The EONIA is the interest rate for the interbank market. It is an important indicator about the current status of European financial markets and the effect of the ECB policy, especially during the financial crisis[18].

Concerning the DAX turnover, we know that our implied volatility measures, VSSE and VDAX, are linked to the DAX variations which in turn are influenced by trading volume (Karpoff, 1986, Karpoff, 1987, Gallant et al., 1992). To measure the impact of the DAX turnover on the long-short portfolios, we add the DAX turnover by volume to the set of control variables. The DAX turnover by volume stands for the total number of constituent shares of the DAX traded at the FSE on a given day.[19]

$$R_{Small,t} - R_{Big,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_t + \beta_E EONIA_{t-1} + \beta_T DAX_{turnover_t} + \varepsilon_t \qquad (2.16)$$

In equation (2.16) the control variables are the lagged EONIA, the DAX turnover by volume, the Fama-French Carhart factors HML, MOM and the market excess return (RMRF). Two control variables include interest rate components, the market excess return (RMRF) and the Sentiment indicator. We test for collinearity following the approach of Belsley et al. (1980, p. 105). The test value is 7.71 meaning that multicollinearity is not an issue if we add EONIA and DAX turnover as control variables. The regression results are presented in table 2.11. The coefficient of the lagged EONIA is positive and significant at the 5 % level (0.000927). The coefficient of the DAX turnover has a negative impact on the return of the long-short portfolio and is significant at the 5 % level. Though these two control variables are significant, they do not change our main result.

[18] The data is available at the German federal bank.
[19] The data is available on Datastream.

Table 2.11 Basic regression controlling for DAX turnover and EONIA

VARIABLES	(1) R_{small}-R_{big}
$Sentiment_{t-1}(x10^4)$	3.19***
	(5.282)
$RMRF_t$	-0.875***
	(-54.15)
HML_t	-0.0194
	(-0.698)
MOM_t	0.0457*
	(1.852)
$DAX\_turnover_t(x10^6)$	-0.00887**
	(-2.504)
$EONIA_{t-1}$	0.000927**
	(2.386)
$Constant(x10^4)$	6.60
	(1.322)
N	1,103
R^2	0.856

Regressing the return of a size-based long-short portfolio on lagged sentiment and Fama-French factors, controlling for DAX turnover and EONIA give the coefficient of sentiment in the simple regression: $R_{Small,t} - R_{Big,t} = \alpha + \beta_S Sentiment_{t-1} + \beta_X X_t + \beta_T Turnover_t + \beta_E EONIA_{t-1} + \varepsilon_t$ controlling for the Fama-French factors, the Carhart momentum factor using the data from Eurofidai, the DAX turnover by volume and the lagged EONIA using data from the German Bundesbank. The sentiment indicator is calculated for the whole sample on a daily basis. The SMB factor is not included as the long-short portfolio is based on this criterion. Without the sentiment indicator the R^2 of the equation is 0.8477. We control for heteroscedasticity by using heteroscedasticity-consistent standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

2.6 Conclusion

Thanks to the unique database of the Stuttgart Stock Exchange (SSE) we are able to introduce a new measure of investor sentiment. We build on the difference in behaviour between retail investors at the SSE and professional investors at the Frankfurt Stock Exchange (FSE). We define an implied volatility index (VSSE) for the SSE to be compared to the implied volatility measures on the DAX at the FSE (VDAX and VDAX-NEW). The comparison allows us to build a sentiment indicator which is significant in predicting the daily returns on a size-based long-short portfolio. As a consequence, our analysis shows the persistent inconsistence between prices of structured products on the SSE and option prices on the FSE.

Our findings are robust to several variations in the definition of the indicator. They are also robust when controlling for a number of factors. In particular, the results remain significant if we calculate different implied volatility measures for the SSE or use another measure for the implied volatility at the FSE. The

use of different types of Fama-French-Carhart factors for the German market supports the idea that our sentiment indicator is robust. Further, we show that the results remain significant when controlling for the autocorrelation of sentiment and variations at the macroeconomic level.

All these results provide empirical evidence that there are significant persistent behavioural differences between the two investor types. Our sentiment indicator partially captures this difference and the persistent mispricing.

Google and Twitter Data: Two Perspectives on the Behaviour of Noise Traders[20]

Carolin Hartmann[21], Hans-Peter Burghof[22] and Marc Mehlhorn[23]

[20] The authors would like to thank Sowa Labs (http://www.sowalabs.com) for providing the sentiment labeled Twitter data about the 30 DJIA companies and the DALAHO of the University of Hohenheim for the data provision (https://wiso.uni-hohenheim.de/en/109608). Further, we want to thank the seminar participants of the Augustin Cournot Doctoral days and the participants of the doctoral seminar of the University of Hohenheim for their comments. Special thanks go to Maxime Merli, Carola Mueller, Monika Gehde-Trapp and Tereza Tykvová for their comments and support. The paper circulted previously under the title "An action – Identifying noise traders entering the market with Google and Twitter"

[21] University of Hohenheim, Institute of Financial Management, Mailing address: Schwerzstraße 38, 70599 Stuttgart, GERMANY, Tel: +49 (0) 7 11 45 92 29 00, Email: burghof@uni-hohenheim.de

[22] University of Hohenheim, Institute of Financial Management, Mailing address: Schwerzstraße 38, 70599 Stuttgart, GERMANY, Tel: +49 (0) 7 11 45 92 29 00, Email: carolin_hartmann@uni-hohenheim.de

[23] TH Köln, Mailing address: Claudiusstraße 1, 50678 Köln, GERMANY, Tel: +49 (0) 221-8275-5159, Email: marc.mehlhorn@th-koeln.de

3 Google and Twitter Data: Two Perspectives on the Behaviour of Noise Traders

3.1 Introduction

Google is a search engine that people use like an online encyclopaedia. Twitter is a microblogging service which people use to tweet, retweet information or to follow other users. Both, Google and Twitter have millions of users worldwide. They are indicators for the public mood (Bollen, Mao, & Zeng, 2011). In July 2018 Twitter had 335 million monthly users[24], 76 % of search engine users choose Google[25]. Especially for financial markets, Twitter becomes more and more important. Former US president Donald Trump used Twitter as his direct medium of communication.[26] Elon Musk's tweet on 7 August 2018 about taking Tesla private led to an increase of the share by 11 %. Short sellers lost approximately 1.3 billion USD.

Research on financial markets shows that GSV and TV have an impact on financial markets (Bollen et al., 2011; Da, Engelberg, & Gao, 2011; Dimpfl & Jank, 2011, 2016; Dimpfl & Kleiman, 2017; Hamid & Heiden, 2015; Kumar & Lee, 2006; Mao, Counts, & Bollen, 2015; Schneller, Heiden, Heiden, & Hamid, 2018).

Google and Twitter capture the behaviour of investors on the market. They offer a huge amount of information but it is questionable if (a) this information is useful and new for financial markets, (b) or used by participants of financial markets, and (c) if Google and Twitter data is reliable.

According to the Efficient Market Hypothesis (Fama, 1981), stock prices should, on average, be driven by fundamentals and only move when new information about fundamentals enter the market and are correctly priced. The idea of behavioural finance is that investors do not behave rational (Kahneman & Tversky, 1979; Shiller, 2003). Instead, we observe investor sentiment and limits-to-arbitrage on financial markets (Baker & Wurgler, 2006, 2007; De Long et al., 1990; Shleifer & Vishny, 1997).

[24] https://de.statista.com/statistik/daten/studie/232401/umfrage/monatlich-aktive-nutzer-von-twitter-weltweit-zeitreihe/
[25] https://de.statista.com/statistik/daten/studie/222849/umfrage/marktanteile-der-suchmaschinen-weltweit/
[26] http://www.trumptwitterarchive.com/

Investor sentiment reflects beliefs about the future of cash flows that are not rationally justified (Baker & Wurgler, 2007). It impacts the decision making of individual investors (See-To & Yang, 2017) and financial market indicators. Researchers try to proxy investor sentiment, as it cannot be measured directly. Barber and Odean (2008) find that abnormal trading volume is the best indirect indicator of investor attention.[27] Tetlock (2007), using media coverage as proxy for investor sentiment, finds that media pessimism can predict a decrease in market prices, leading the return back to fundamentals. Da et al. (2011) are among the first to use the GSV to proxy investor attention to predict stock price movements. They find that an increase in investor attention can lead to a stronger sentiment, especially when attention comes from noise traders. Further authors confirm the predictive power of GSV on financial market indicators (Dimpfl & Jank, 2016; Dimpfl & Kleiman, 2017; Fink & Johann, 2014). With respect to stock market volatility, Dimpfl and Jank (2011) find that an increase in GSV increases the volatility of the DJIA on the next day.[28] Hamid and Heiden (2015) find a significantly better in-sample and out-of-sample predictions including GSV, especially in times of high volatility. These findings are in line with the outcome of further studies (Andrei & Hasler, 2015; Choi & Varian, 2012; Vlastakis & Markellos, 2012; Vozlyublennaia, 2014).

The microblogging platform Twitter unites several functions, such as tweeting, retweeting, following and reading. Bollen et al. (2011) find that moods extracted from Twitter tweets are able to improve DJIA predictions. Alexander and Gentry (2014) show that 77 % of the Fortune 500 companies in America tweeted in 2013. They point out that companies use social media platform to republish company information and to have a channel to keep investors updated e.g. for live tweeting during special events such as annual general meeting (SEC, 2013). Tafti et al. (2016) find a real-time relationship between the activity on Twitter and the trading volume of Nasdaq100 firms. Sprenger et al. (2014) point out that microblogs possess valuable information that is not captured by market indicators yet. Mao et al. (2015) find that the bullishness of Twitter updates on a daily level is a suitable indicator for investor sentiment. They show that there is a positive correlation between Twitter bullishness and Google. Changes in the

[27] Investor attention measure to what investors pay attention to (Da et al., 2011). This is important as attention of individuals is limited (Kahneman, 1973).
[28] As a measure of volatility, they use realized volatility which they calculate following Andersen et al. (Andersen et al., 2003).

bullishness of Twitter are followed by changes in Google, suggesting a lead-lag relationship between the two. See-To and Yang (2017) confirm that Twitter is a direct measure of investor sentiment.

In this chapter we use GSV and TV. Therefore we have a unique daily dataset with data from Google Trends and Sow Labs on the DJIA. We assume that Google and Twitter data contain new information on the DJIA. They reflect the behaviour of individual investors that is not yet considered by other market indicators. The first hypothesis is that GSV and TV have an impact on the trading activity leading to more traders on the market. In line with the literature (Easley et al., 1996) we find that changes in GSV and TV have an impact on turnover on the same day and the next day.

The second hypothesis is that an increase in TV and GSV lead to an increase in the share of noise traders on the market. Based on the DSSW (De Long et al., 1990), we find that GSV and TV are positively correlated with volatility of the DJIA. Further, TV leads to an increase in volatility on the same day and the next day. We assume that TV captures the movements on financial markets better than Google. While people use Twitter to share news, Google is a search engine where people Google things that are already in the news. Due to the lead-lag relationship mentioned in the literature (Mao et al., 2015).

This chapter is organized as follows. Section 2 outlines our methodology. Section 3 describes our dataset. Section 4 presents the empirical results and assess the impact of GSV and TV on (1) stock turnover following the model of Easley et al. (1996) and (2) volatility following the DSSW model. Section 5 shows several robustness test. Section 6 looks at the possible shortcomings of online data. Section 7 concludes.

3.2 Methodology

3.2.1 Trading activity

We use a two-step procedure to measure (1) the impact of new information on trading and (2) the change in the share of noise traders on the market due to changes in GSV and TV. As a first step we measure the arrival of new traders on the market. We follow the idea of Mehlhorn (2018) and adapt an approach used by Easley et al. (1996) and Fink and Johann (2014). Their market microstructure model measures

the probability of informed trading. According to Easley et al. (1996), the probability of informed trading is lower (higher) for high (low) volume stocks. It depends on new information on the market. The existence of new information increases the number of traders on the markets. Normally informed and uninformed traders are on the market.[29]

We simplify the model approach to test the implication of GSV and TV on the trading activity on the market.[30] We use GSV and TV as a proxy for new information. Second, we measure if GSV and TV have an impact on trading activity. We consider all stock that are part of the DJIA.[31] Furthermore, we look at turnover instead of volume. We measure turnover as the ratio between the amount of shares traded and the number of shares outstanding (Lo & Wang, 2000). This allows us to qualify for different trading volumes and number of shares outstanding of stocks. We expect to find an impact of changes in GSV and TV on trading, if they include new information.

In a second step we assess if GSV and TV increase the share of noise traders on the market. The idea dates back to Black (1986). He finds that noise trading can create volatility in the future. Noise is a source of inefficiency which make trading on financial markets possible. Foucault et al. (2011) look at the effect of a French stock market reform which increases costs for speculative trading for individual investors. Applying the DSSW model (De Long et al. 1990), they find that changes in volatility are not entirely explained by changes of the fundamental value or changes in news. Individual investors who behave like noise traders increase volatility.

We follow the formal approach of the DSSW model of DeLong et al. (1990). It distinguishes between sophisticated investors, who are rational (institutional) traders, and noise (individual) traders, who are irrational traders. Noise traders are subject to sentiment, meaning that their beliefs about future prices deviate from the fundamental value (Long et al. 1990; Baker & Wurgler 2007). The unpredictable behaviour of noise traders increases the trading risk for rational traders (De Long et al., 1990). Rational investors cannot foresee the reaction of the noise traders. Thus, it takes some time until rational investors

[29] Informed traders are risk neutral. The trading of uninformed traders is noise. New information arrive on the market with a certain probability.
[30] We have no personal data, we do not know if really more single trader enter the market or if the traders on the market trade more.
[31] In the initially model, Easley et al. (1996) look at over 1000 stocks. Here, the differences between the stocks and the trading volumes are more pronounced. As the 30 stocks of the DJIA are all blue chip stocks, we do not differentiate between high and low volume stocks.

and arbitrageurs force prices back to their fundamental value (Baker & Wurgler, 2007; Barberis & Thaler, 2003; De Long et al., 1990; Hirshleifer, 2001; Shleifer & Vishny, 1997).

As we cannot measure noise traders directly, we proxy investor attention and sentiment using Google and Twitter data to model the behaviour of individual investors (Antweiler & Frank, 2004; Audrino, Sigrist, & Ballinari, 2020; Bollen et al., 2011; Da et al., 2011; Hamid & Heiden, 2015; Kumar & Lee, 2006; Mao et al., 2015). The more noise traders on the market, the higher is volatility. Antweiler and Frank (2011) assess internet stock messages on the DJIA and find that they help to predict market volatility. If individual investors behave like noise traders, this can have a positive effect on volatility, contributing to idiosyncratic volatility above and beyond cash-flow news (Foucault et al. 2011).

3.2.2 Noise traders in the DSSW model

The DSSW model (De Long et al., 1990) is a two-period overlapping generations model. The aim of the agents is to maximize their utility. There are two types of agents. The sophisticated investors i and the noise traders n. The share of noise traders is equal to μ; the share of sophisticated investors is equal to $1-\mu$. In the first period t, they can decide between an investment in a safe asset s and in a risky asset u. The price of the safe asset is equal to one. The price of the risky asset is equal to p_t in period t. The price p_{t+1} in the second period is unknown. It depends on the agents' expectations. The expectations of the sophisticated investors are based on fundamentals. Noise traders are subject to sentiment as their beliefs about the price development are misperceived. The misperception is measured by ρ_t. The risk aversion of the agents is denoted by the coefficient γ. In the DSSW model, the price of the risky asset today p_t is a discounted function of the expected price of the risky asset in $t+1$:

$$p_t = \frac{1}{1+r}\left[r + {}_tp_{t+1} + \mu\rho_t - 2\gamma\ {}_t\sigma^2_{p_{t+1}}\right]. \tag{3.1}$$

The price p_t consists of the return the agents will earn in the second period r; the expected price of the risky asset in the second period ${}_tp_{t+1}$, the misspecification of the price by the noise trader times the share of the noise traders on the market $\mu\rho_t$, the behaviour of the agents according to risk γ and the variance of the risky asset ${}_t\sigma^2_{p_{t+1}}$. Assuming a steady state equilibrium the price of the risky asset p_t

depends on known parameters. Based on these assumptions of the DSSW (De Long et al., 1990), we look at the effect of the share of noise traders on volatility. According to the DSSW model, the variance of the future price is defined as follows:

$$\sigma^2{}_{p_{t+1}} = \frac{\mu^2\sigma_\rho^2}{(1+r)^2}. \tag{3.2}$$

To determine the influence of noise traders on the variance of the price we take the partial differential of the variance $\sigma^2_{p_{t+1}}$ with respect to the share of the noise traders μ:

$$\frac{\partial\sigma^2_{p_{t+1}}}{\partial\mu} = \frac{1}{(1+r)^2}\ 2\mu\sigma_\rho^2 > 0. \tag{3.3}$$

The share of noise traders has a positive effect on the variance of the price of the risky asset tomorrow. If the share μ of noise traders on the market increases, the variance of the risky price in the next period increases as well. The size of the effect is determined by the share of noise traders on the market, the variance of the bullishness σ_ρ^2 of the noise traders and the quadratic discount factor $(1+r)^2$. From the effect of noise traders on the variance, we can deduce the effect on the volatility. An increase in the share of noise traders μ leads to an increase in the variance and therefore in the standard deviation σ. The more noise traders are on the market, the higher the volatility on the market and *vice versa*.

3.3 Data and descriptive statistics

3.3.1 Dataset and sample

The sample we use is unique. It consists of GSV data from Google Trends and as a new source TV from Sowa Labs. We only consider stock market related topics and cover a time period from 6 June 2013 until 31 December 2016. The balanced panel contains market data for 29 stocks of the DJIA from Datastream[32]. A list of the stocks can be found in appendix 1. One observation stands for one stock per

[32] Trading data on the DJIA comes from NYSE and NASDAQ.

day. We include only weekdays from Monday to Friday to avoid unclear weekend effects (Andersen, Bollerslev, Diebold, & Labys, 2003).[33] All variables are expressed as percentage changes.

Google

The data on GSV is obtained via Google Trends (https://trends.google.de/trends/) as shown in earlier research (Dimpfl & Jank, 2016; Fink & Johann, 2014; Hamid & Heiden, 2015). This specific Google website gives the possibility to obtain information on the search volume for various search terms. As a result, the GSV Index is generated. GSV is a time series which depict the relative search volume of a specific search term on Google during a certain time span or point in time. It can take the value from 0 to 100. Depending on the length of the time span, the time series ranges from eight minutes intervals up to monthly data. It is further possible to create filters, for example for regions or special topics such as "all categories" or "finance". Moreover, there is a Google trend feature which allows to specify the search term. In addition the option "search term", it is possible to choose a "topic". This option covers a broader range for a specific search term than the search term alone. For the search of the constituents of the DJIA index, the finance filter in combination with the topic filter was chosen on a daily level. The data is obtained for 29 stocks being part of the DJIA in this time span. The variable ΔGSV represents the daily percentage changes in GSV. As we have panel data, we measure the percentage changes for each stock of the DJIA at time t for company i.[34]

$$\Delta GSV_{i,t} = ln\left(\frac{GSV_{i,t}}{GSV_{i,t-1}}\right). \tag{3.4}$$

Twitter

The Twitter data is obtained from Sowa Labs (https://www.sowalabs.com/) and comparable to the data used by Peter et al. (Peter, Darko, Igor, & Miha, 2017). It contains approximately 4.5 million tweets about 29 companies. The data is collected by a Twitter search Application Programming Interface (API). For each company, the stock cashtag is specified (e.g. "$AAPL" refers to Apple). To capture the market movement, we use the TV on each stock without distinguishing between positive, negative or neutral

[33] We exclude the GSV and TV on Saturdays, Sundays and holidays to avoid interference with days of lower trading activity or none availability of data.
[34] The use of logarithmic variables leads to better behaved variables and reduces the correlation between the independent variables.

sentiment. The tweets are aggregated on a daily level. We use the daily percentage changes in TV denoted as ΔTV.

Turnover

As a measure of trading activity we use turnover. The advantage of turnover over trading volume is that turnover allows to control for the size of the company. Looking at turnover as a proportion between trading volume and number of shares outstanding avoids overweighting of large stocks. Lo and Wang (2000) calculate turnover as a share between the volume of traded shares and the shares outstanding.[35] Brooks (1998) takes the shares traded per day divided by the number of shares outstanding. Trading volume and number of shares outstanding are from Datastream. Turnover is calculated for each stock on a daily basis. Our variable is the daily percentage change in turnover $\Delta Turnover$.

Volatility

Volatility allows us to capture the movement of the stock markets. We use an approach based on historical data suggested by Brooks (Brooks, 1998, 2014). We measure volatility as the logarithmic difference between high and low stock prices per day for each stock following equation 3.5. The daily high and low prices are from Datastream. Our volatility variable is the daily percentage changes in Volatility $\Delta Volatility$.

$$Volatility_{i,t} = ln\left(\frac{Price\ high_{i,t}}{Price\ low_{i,t}}\right). \tag{3.5}$$

Squared returns

As an indicator for news on the market, we use squared returns to capture larger market movements. Extreme returns can be an indicator for news Barber and Odean (2008) but according to Andersen and Bollerslev (1998), squared daily returns can also be seen as a noisy proxy for the true volatility. Abhyankar (1995) measures good or bad news by the size of returns (Brooks, 1998). Vozlyublennaia (2014) uses past returns as a measure of information. In line with the literature we use squared returns as a proxy for the impact of news on the market. The continuous returns are calculated based on daily prices

[35] Turnover is calculated as: $Turnover_{it} = \frac{X_{i,t}}{N_i}$, with $X_{i,t}$ being the volume of shares traded of one stock i at time t and N_i standing for the total number of shares outstanding of stock i.

for all relevant stocks. The data is from Datastream. Our variable reflects the daily changes in squared returns $\Delta Squared\_Returns$.

3.3.2 Summary statistics

Table 3.1 presents the summary statistics. The mean value of all variables is close to zero. The standard deviation varies between 0.323 (ΔGSV) and 2.992 ($\Delta Volatility$). Fat tails are more pronounced for changes in TV (0.283) than for changes in GSV (0.080). For changes in GSV (TV), we find a positive skewness 0.080 (0.283). The tail on the right is longer than on the left. The kurtosis for changes in GSV (TV) is 6.059 (6.173). The changes in turnover are positively skewed (0.28) and peaked (4.91). This means that there are more extreme changes in turnover than under normal distribution and they are more pronounced on the right, as the skewness is larger than zero. The distribution of changes in volatility and squared returns is close to normal. The number of observations varies as some observations are equal to zero and here we cannot calculate the changes.

Table 3.1 Summary statistics

VARIABLES	Mean	S.D.	Skewness	Kurtosis	Min	Max	N
ΔGSV	0.0001571	0.3230409	0.0803382	6.059349	-2.079442	2.628801	25544
ΔTV	0.0000635	0.5802832	0.2832056	6.17311	-4.382027	4.567814	26136
ΔSquared_Return	-0.0006845	0.4739944	0.1554362	3.219353	-2.311268	2.507094	26158
ΔTurnover	-0.0005657	0.3484734	0.2756585	4.910202	-1.65539	2.576512	26158
ΔVolatility	-0.0103956	2.99265	-0.0186688	3.597933	-13.40264	12.78078	25781

ΔGSV and ΔTV stand for percentage changes in GSV and TV. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. ΔVolatility stands for the percentage changes in volatility. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA.

3.4 Empirical study

3.4.1 Correlation analysis

The correlation presented in table 3.2 are below 0.3 for almost all variables. The correlation between changes in turnover and changes in volatility is 0.4967. This coefficient is high and significant at all reasonable significance levels meaning that volatility and turnover increase together. Brooks et al. (Brooks, Rew, & Ritson, 2001) confirm this positive relationship. The correlation between changes in

squared returns and volatility is 0.4358. Changes in GSV (TV) are positively correlated with changes in turnover and volatility. The correlation is higher for changes in TV. Thus, we expect the effect of changes in TV to be more pronounced than the effect of changes in GSV on turnover and volatility. This is in line with Mao et al. (2015) who find that Twitter is a better predictor of investor sentiment than Google. Moreover, changes in GSV and TV are positively correlated (0.117). To assess multicollinearity, we calculate the condition number. In our sample, the condition number is equal to 1.68, which is below the critical threshold of 20 (D.A. Belsley et al., 1980, p. 58; David A. Belsley, 1991). Therefore, there is no multicollinearity problem in our sample.

Table 3.2 Correlation analysis

	ΔSquared_Return	ΔVolatility	ΔGSV	ΔTV	ΔTurnover
ΔSquared_Return	1.0000				
ΔVolatility	0.4358***	1.0000			
ΔGSV	0.0441***	0.0674***	1.0000		
ΔTV	0.1355***	0.2034***	0.1166***	1.0000	
ΔTurnover	0.3301***	0.4967***	0.1165***	0.2891***	1.0000

ΔGSV and ΔTV stand for percentage changes in GSV and TV. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. ΔVolatility stands for the percentage changes in volatility. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

3.4.2 Panel data regressions

To analyse the impact of changes in GSV and TV on turnover and volatility, we implement the models of Easley et al. (1996) and DeLong et al. (Long et al., 1990). Therefore, we apply a panel fixed effects model with firm and time fixed effects. The estimation equation follows the general form (Brooks, 2014):

$$y_{i,t} = \alpha + \beta x_{i,t} + \mu_i + \lambda_t + v_{i,t}, \tag{3.6}$$

where y_{it} is the dependent variable at time t for company i. The variable α stands for the time invariant intercept. The coefficient β is a $kx1$ vector and the independent variable x_{it} is a $1xk$ vector. The time fixed effect is λ_t and the cross-sectional fixed effect μ_i. Time $t = 1, \dots T$ is a daily measure and $i = 1, \dots, N$ represents each company which is part of the sample.

The procedure is twofold. First, we measure the impact of changes in GSV and TV on trading activity. The dependent variable in equation (3.7) is the change in turnover ($\Delta Turnover_{i,t}$). Following the idea of Easley et al. (1996) new information leads to new market entries of traders. Hence, we expect that an increase in turnover leads to a rising number of traders arriving on the market. We use GSV and TV as proxies for new information. We expect that an increase (decrease) in GSV and TV leads to an increase (decrease) in turnover. We quantify this effect by looking at the percentage change in turnover. The control variables are lagged changes in turnover, changes in squared returns and lagged squared returns. Step by step, we include changes in GSV (TV) and lagged changes in GSV (TV) to obtain the final equation:

$$\Delta Turnover_{i,t} = \alpha + \beta_1 \Delta Turnover_{i,t-1} + \beta_2 \Delta SquaredReturn_{it} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta GSV_{i,t} + \beta_5 \Delta GSV_{i,t-1} + \beta_6 \Delta TV_{i,t} + \beta_7 \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}. \quad (3.7)$$

Second, we apply the DSSW model (De Long et al., 1990) to measure the impact of changes in GSV and TV on the amount of noise traders on the market. The dependent variable is changes in volatility. A positive (negative) coefficient leads to an increase (decrease) in volatility. According to the DSSW model (De Long et al., 1990), this means that the share of noise traders on the market increases (decreases). We control for changes in lagged volatility, squared returns and lagged squared returns. Step by step we include changes in GSV (TV) and lagged changes in GSV (TV).

$$\Delta Volatility_{i,t} = \alpha + \beta_1 \Delta Volatility_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta Turnover_{i,t-1} + \beta_5 \Delta Turnover_{i,t-1} + \beta_6 \Delta GSV_{i,t} + \beta_7 \Delta GSV_{i,t-1} + \beta_8 \Delta TV_{i,t} + \beta_9 \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}. \quad (3.8)$$

Table 3.3 displays the results of the fixed effects regression (8) on changes in turnover. All coefficients are highly significant at the 1 % level. In column (1) a baseline of control variables is included (lagged changes in turnover, changes in squared returns and lagged changes in squared returns). Column (2) includes changes in GSV, column (3) changes in lagged GSV. Column (4) includes changes in TV, column (5) includes changes in lagged TV. R^2 increases from 49 % in column (1) to 56 % in column (5). In all five equations, we implement time and company fixed effects. To control for cross sectional

dependence, heteroscedasticity and autocorrelation, we use Driscoll and Kraay standard errors (Driscoll & Kraay, 1998).

Our results show that days with high trading activity are followed by days with a decrease in trading activity. A positive change in turnover by 10 % in $t-1$, leads to a decrease in turnover in t by -2.96 %. On the same day, a 10 % change in squared returns (lagged squared returns) leads to an increase of changes in turnover around 0.3 % (0.2 %). We find that, controlling for changes in squared returns and lagged turnover, GSV and TV have a positive and highly significant impact on turnover. An increase in changes of GSV (TV) by 10 % leads to an increase in turnover on the same day by 0.51 % (1.73 %). An increase in lagged changes of GSV (TV) by 10 % leads to an increase in turnover today by 0.27 % (0.92 %). In the sense of Easley et al. (1996), GSV and TV incorporate news that are new to the market. As a consequence trading activity increases and more traders enter the market.

Further, the effect of TV is higher than the effect of GSV. This supports the finding of Mao et al. (Mao et al., 2015) that Google and Twitter have a lead-lag structure. This is in line with the correlation we find in Table 3.2 and the study of Mao et al. (2015). We conclude that TV is of high relevance for the trading activity of investors.

Table 3.3 Market entry – basic approach

VARIABLES	(1) $\Delta Turnover_t$	(2) $\Delta Turnover_t$	(3) $\Delta Turnover_t$	(4) $\Delta Turnover_t$	(5) $\Delta Turnover_t$
$\Delta Turnover_{i,t-1}$	-0.296***	-0.294***	-0.299***	-0.287***	-0.341***
	(-36.874)	(-36.253)	(-38.043)	(-35.482)	(-50.660)
$\Delta Squared\_Return_{i,t}$	0.038***	0.038***	0.038***	0.033***	0.031***
	(30.039)	(30.508)	(30.712)	(32.701)	(32.837)
$\Delta Squared\_Return_{i,t-1}$	0.020***	0.020***	0.019***	0.018***	0.016***
	(22.976)	(22.933)	(22.802)	(23.756)	(22.916)
$\Delta GSV_{i,t}$		0.071***	0.095***	0.066***	0.051***
		(8.467)	(9.152)	(8.011)	(6.627)
$\Delta GSV_{i,t-1}$			0.049***	0.045***	0.027***
			(6.458)	(6.316)	(4.068)
$\Delta TV_{i,t}$				0.147***	0.173***
				(23.162)	(24.356)
$\Delta TV_{i,t-1}$					0.092***
					(18.783)
Constant	0.060***	0.049***	0.047***	-0.051***	0.023***
	(38.549)	(21.977)	(18.724)	(-10.252)	(6.232)
Time Fixed Effects	YES	YES	YES	YES	YES
Company Fixed Effects	YES	YES	YES	YES	YES
R^2	0.492	0.495	0.497	0.545	0.560
N	25,566	24,970	24,833	24,811	24,801

Column (1) to (5) report the result on market entry:$\Delta Turnover_{i,t} = \alpha + \beta_1 \Delta Turnover_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta GSV_{i,t} + \beta_5 \Delta GSV_{i,t-1} + \beta_6 \Delta TV_{i,t} + \beta_7 \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}$. Column (1) includes the control variables changes in lagged turnover, changes in squared returns and changes in lagged squared returns. In column (2) the variable change in GSV is added. Column (3) adds the lagged change in GSV. Column (4) adds the change in TV. Column (5) adds the lagged change in TV. The dependent variable ΔVolatility stands for the percentage changes in volatility. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. All regressions are panel fixed effects regression, including time and company fixed effects. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

Table 3.4 represents the result of the panel fixed effect regression on changes in volatility (8). Column (1) includes the control variables changes in lagged volatility, turnover, lagged turnover, squared returns and lagged squared returns. Column (2) includes the changes in GSV and column (3) the lagged changes in GSV. Column (4) includes the changes in TV and column (5) the changes in lagged TV. An increase in yesterday's volatility by 10 % would lead to a decrease in volatility today by -4.74 %. We observe a mean reversion trend for volatility. As expected by the correlation, the coefficients of changes in turnover and lagged turnover are always positive and highly significant at the 1 % level. This means that an increase in trading activity leads to a higher volatility on the market. An increase in turnover today (yesterday) by 10 % would lead to an increase in volatility by 5.71 % (2.5 %). The same holds for news

proxied by squared returns. An increase in squared returns on same day by 10 % would lead to an increase in volatility by 0.47 %. The impact from the previous day is smaller with 0.22 %.

In column (2) to (5) we include GSV and TV to proxy for investor sentiment and investor attention. We find that only the effect of changes TV are highly statistically significant at the 1 % level. A 10 % increase in TV would on the same day (day before) lead to an increase in volatility by 0.4 % (0.2 %). This means that TV incorporates new information for the market which is not captured by the other control variables so far. With regard to the DSSW model we can say that an increase in TV is an indicator for the increase of the share of noise traders on the market. Changes in GSV and lagged GSV are not significant meaning that they have no impact changes in volatility.

Like Mao et al. (2015), we find that TV is more important for financial markets than GSV. With TV we measure new information from investors that comes on the market. The share of noise traders increase (decreases) if TV increases (decreases). We expect that this deviation from the mean is only temporary (Mao et al., 2015). Arbitrageurs and rational investors bring the changes in volatility back to zero but as they face the risk to trade against noise traders it takes a while (De Long et al., 1990; Shleifer & Vishny, 1997).

Table 3.4 Share of noise trader – basic model

VARIABLES	(1) $\Delta Volatility_t$	(2) $\Delta Volatility_t$	(3) $Volatility_t$	(4) $Volatility_t$	(5) $Volatility_t$
$\Delta Volatility_{i,t-1}$	-0.474***	-0.474***	-0.474***	-0.473***	-0.475***
	(-84.207)	(-82.124)	(-82.114)	(-82.277)	(-82.730)
$\Delta Turnover_{i,t}$	0.571***	0.574***	0.574***	0.550***	0.543***
	(44.500)	(44.194)	(44.198)	(41.829)	(40.157)
$\Delta Turnover_{i,t-1}$	0.250***	0.250***	0.250***	0.245***	0.232***
	(26.312)	(25.718)	(25.435)	(25.257)	(21.979)
$\Delta Squared\_Return_{i,t}$	0.047***	0.047***	0.047***	0.047***	0.047***
	(35.853)	(35.259)	(35.622)	(35.424)	(35.283)
$\Delta Squared\_Return_{i,t-1}$	0.022***	0.023***	0.022***	0.022***	0.022***
	(22.541)	(22.030)	(22.047)	(22.186)	(22.020)
$\Delta GSV_{i,t}$		-0.004	-0.002	-0.007	-0.010
		(-0.616)	(-0.243)	(-0.862)	(-1.213)
$\Delta GSV_{i,t-1}$			0.004	0.004	0.001
			(0.548)	(0.525)	(0.069)
$\Delta TV_{i,t}$				0.036***	0.043***
				(7.560)	(8.386)
$\Delta TV_{i,t-1}$					0.020***
					(3.720)
Constant	0.106***	0.106***	0.106***	-0.136***	-0.024***
	(53.016)	(45.415)	(44.074)	(-42.267)	(-4.053)
Time Fixed Effects	YES	YES	YES	YES	YES
Company Fixed Effects	YES	YES	YES	YES	YES
R^2	0.593	0.593	0.594	0.595	0.596
N	25,566	24,970	24,833	24,811	24,801

Column (1) to (5) report the result on market entry: $\Delta Volatility_{i,t} = \alpha + \beta_1 \Delta Volatility_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta Turnover_{i,t} + \beta_5 \Delta Turnover_{i,t-1} + \beta_6 \Delta GSV_{i,t} + \beta_7 \Delta GSV_{i,t-1} + \beta_8 \Delta TV_{i,t} + \beta_9 \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}$. Column (1) includes the control variables changes in lagged turnover, changes in squared returns and changes in lagged squared returns. In Column (2), the variable GSV is added. Column (3) adds the lagged change in GSV. Column (4) adds the change in TV. Column (5) adds the lagged changes in TV. The dependent variable ΔVolatility stands for the percentage changes in volatility. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. All regressions are panel fixed effects regression, including time and company fixed effects. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

We are one of the first to compare the influence of changes in GSV and TV in a panel data setting for the DJIA. We find that GSV and TV have a positive impact on changes in turnover. In the sense of Easley et al. (1996), we find that information in GSV and TV lead to market entries. The effect is more pronounced for TV than for GSV. Furthermore, we find that both have predictive power on turnover. The coefficients are smaller but still significant from one day to the next.

For volatility, we find a significant impact of changes in TV on volatility, while GSV has no impact. In the sense of De Long et al. (Long et al., 1990), we find that TV leads to an increase of the share of noise

traders on the market. This effect is observable for the same day and the next, indicating the predictive power of TV on volatility.

Twitter seems to observe the financial markets better than Google. If someone tweets or retweets a statement concerning a share of a DJIA index company there is a better chance that it translates into a trade. Another reason could be the lead lag relationship between Google and Twitter (Mao et al., 2015). People use Google to verify the Twitter news which can explain why the influence of Google gets lower as soon as Twitter is integrated in the search. We do not expect that people use Google before they get new information via Twitter.

3.5 Robustness tests

3.5.1 Model fit

For the regressions in part 3.4 we use a panel data model with time fixed effects, company fixed effects and Discroll and Kraay standard errors. Our results show that we have cross sectional dependence. This means that the residuals are correlated across the different companies i in the panel. To overcome this problem, we apply Driscoll and Kraay standard errors (Driscoll & Kraay, 1998)[36]. Further, we find that we have heteroscedasticity meaning that our variance is not constant. We can solve this problem by using robust standard errors[37]. Here we use Huber and White robust standard errors, or the Driscoll and Kraay standard errors. We run our estimations with Driscoll and Kraay, Huber and White and clustered standard errors. Overall, the coefficients and the significance levels do not change, except for the constant.[38]

In Table 3.5 we compare different ways to estimate the standard equation (3.7) for turnover including all control variables. In column (1), we use Huber and white standard errors. In column (2), we use Driscoll and Kraay standard errors. We find that our coefficients do not change. The standard errors in

[36] See Hoechle (2007)
[37] For the robust standard errors see Huber (1967) and White (White, 1980)
[38] We do not report the results of the clustered standard errors on the company level, as they lead to the same results than the Huber and White standard errors.

column for Huber and White in column (1) are smaller than for Driscoll and Kraay. The main difference is the constant, which is larger for the robust standard errors in column (1).

Table 3.5 Market entry – different standard errors

VARIABLES	(1) $\Delta Turnover_t$	(2) $\Delta Turnover_t$
$\Delta Turnover_{i,t-1}$	-0.341***	-0.341***
	(-58.173)	(-50.660)
$\Delta Squared\_Return_{i,t}$	0.031***	0.031***
	(27.549)	(32.837)
$\Delta Squared\_Return_{i,t-1}$	0.016***	0.016***
	(18.389)	(22.916)
$\Delta GSV_{i,t}$	0.051***	0.051***
	(3.998)	(6.627)
$\Delta GSV_{i,t-1}$	0.027**	0.027***
	(3.231)	(4.068)
$\Delta TV_{i,t}$	0.173***	0.173***
	(20.317)	(24.356)
$\Delta TV_{i,t-1}$	0.092***	0.092***
	(11.838)	(18.783)
Constant	0.144**	0.023***
	(3.445)	(6.232)
Time Fixed effects	YES	YES
Company Fixed effects	YES	YES
R-squared	0.560	0.560
N	24,801	24,801

Column (1) and (2) report the result on market entry: $\Delta Turnover_{i,t} = \alpha + \beta_1 \Delta Turnover_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta GSV_{i,t} + \beta_5 \Delta GSV_{i,t-1} + \beta_6 \Delta TV_{i,t} + \beta_7 \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}$. In column (1) we use Huber and White robust standard errors. In column (2) we use Driscoll and Kraay standard errors. . All regressions are panel fixed effects regression, including time and company fixed effects. The dependent variable ΔTurnover stands for the percentage changes in trading activity. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

In Table 3.6, we perform regression (8) on changes in volatility including all control variables. The coefficients do not change, only the constant changes. In column (1), we use Huber and White standard errors. In column (2), we use Driscoll and Kraay standard errors. For lagged changes, the significance level differs, as the standard error is higher in column (2) than in column (1). The constant is positive for the Huber and White standard error and close to zero but negative for the Driscoll and Kraay standard error.

Table 3.6 Share of noise trader – different standard errors

VARIABLES	(1) $\Delta Volatility_t$	(2) $\Delta Volatility_t$
$\Delta Volatility_{i,t-1}$	-0.475***	-0.475***
	(-88.948)	(-82.730)
$\Delta Turnover_{i,t}$	0.543***	0.543***
	(31.395)	(40.157)
$\Delta Turnover_{i,t-1}$	0.232***	0.232***
	(20.263)	(21.979)
$\Delta Squared\_Return_{i,t}$	0.047***	0.047***
	(29.859)	(35.283)
$\Delta Squared\_Return_{i,t-1}$	0.022***	0.022***
	(23.252)	(22.020)
$\Delta GSV_{i,t}$	-0.010	-0.010
	(-1.093)	(-1.213)
$\Delta GSV_{i,t-1}$	0.001	0.001
	(0.066)	(0.069)
$\Delta TV_{i,t}$	0.043***	0.043***
	(7.180)	(8.386)
$\Delta TV_{i,t-1}$	0.020**	0.020***
	(3.511)	(3.720)
Constant	0.241***	-0.024***
	(5.156)	(-4.053)
Time Fixed effects	YES	YES
Company Fixed effects	YES	YES
R-squared	0.596	0.596
N	24,801	24,801

Column (1) and (2) report the result on market entry: $\Delta Volatility_{i,t} = \alpha + \beta_1 \Delta Volatility_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta BidAsk_{i,t} + \beta_5 \Delta BidAsk_{i,t-1} + \beta_6 \Delta Turnover_{i,t} + \beta_7 \Delta Turnover_{i,t-1} + \beta_8 \Delta GSV_{i,t} + \beta_9 \Delta GSV_{i,t-1} + \beta_{10} \Delta TV_{i,t} + \beta_{11} \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}$. In column (1) we use Huber and White robust standard errors. In column (2) we use Driscoll and Kraay standard errors. The dependent variable ΔVolatility stands for the percentage changes in volatility. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. All regressions are panel fixed effects regression, including time and company fixed effects. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

3.5.2 Market liquidity

The impact of TV on volatility remains significant at the 1 % level if we include spreads as a measure for market liquidity in the model. We measure the liquidity following Corwin and Schulz (2012). The original approach is a bid-ask spread estimator from daily high and low prices.[39] The idea behind that estimator is that the high-low ratio reflects both the stock's variance and its bid-ask spread. The market

[39] An example to calculate the bid ask spread is provided on Corwin's homepage https://www3.nd.edu/~scorwin/

is more (less) liquid on days with low (high) spreads. We calculate the spread measure following Corwin and Schulz (2012) and compute the percentage change of the spread.[40]

Table 3.7 displays the summary statistic on spreads. In Panel A, we find that the mean change is close to zero. In Panel B, we report the correlation between changes in spreads and other variables. An increase in spreads is negatively correlated with changes in squared returns, volatility, GSV, TV and turnover. For turnover, this means that an increase in the spread, thus a wider bid-ask spread, is correlated with a decrease in turnover. Wider spreads lead to less liquid markets which leads to a decrease in turnover and volatility. We also look at the spread of each stock in the sample and find that the mean is always close to zero. The table can be found in appendix 2.

Table 3.7 Descriptive statistic on spreads

Panel A: Summary Statistics							
VARIABLES	Mean	S.D.	Skewness	Kurtosis	Min	Max	N
ΔSpread	0.0258852	1.2613	0.0012	5.8008	-8.5027	9.6582	16076
Panel B: Correlation Analysis							
VARIABLES	ΔSquared_Return	ΔVolatility	GSV	TV	Turnover	Spread	
ΔSpread	-0.0748***	-0.2536***	-0.0171	-0.0298***	-0.0218***	1.000	

Panel A reports the summary statistic and Panel B the correlation analysis of the changes in spreads. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels respectively.

In Table 3.8 we include the changes in spreads in regression (8) on volatility:

$$\Delta Volatility_{i,t} = \alpha + \beta_1 \Delta Volatility_{i,t-1} + \beta_2 \Delta \text{SquaredReturn}_{i,t} + \beta_3 \Delta \text{SquaredReturn}_{i,t-1} +$$
$$+ \beta_4 \Delta \text{Turnover}_{i,t-1} + \beta_5 \Delta \text{Turnover}_{i,t-1} + \beta_6 \text{Spread}_{i,t} + \beta_7 \text{Spread}_{i,t-1} + \beta_8 \Delta \text{GSV}_{i,t} +$$
$$\beta_9 \Delta \text{GSV}_{i,t-1} + \beta_{10} \Delta \text{TV}_{i,t} + \beta_{11} \Delta \text{TV}_{i,t-1} + \mu_i + \lambda_t + v_{i,t}. \quad (3.9)$$

Including spreads for market liquidity increases the goodness of fit of the model (R^2) to 63 %.We find that the impact of market liquidity on volatility is negative and significant at the 1% level. An increase in spreads by 10% has a negative impact on volatility (-0.57%). If spreads increase, markets are less liquid and less volatile. The number of observations decreases to around 10,000 as we have more missing observations for spreads.

[40] We also compute the zero spread but due to a low number of observations, we do not consider it here.

Table 3.8 Share of noise trader – market liquidity

	(1) $\Delta Volatility_t$	(2) $\Delta Volatility_t$	(3) $\Delta Volatility_t$	(4) $\Delta Volatility_t$	(5) $\Delta Volatility_t$
$\Delta Volatility_{i,t-1}$	-0.440***	-0.440***	-0.440***	-0.439***	-0.442***
	(-36.924)	(-36.270)	(-36.039)	(-35.792)	(-36.626)
$\Delta Turnover_{i,t}$	0.538***	0.541***	0.543***	0.518***	0.506***
	(27.613)	(27.060)	(26.933)	(24.774)	(23.741)
$\Delta Turnover_{i,t-1}$	0.229***	0.231***	0.233***	0.228***	0.208***
	(14.500)	(14.551)	(14.521)	(14.155)	(12.001)
$\Delta Squared\_Return_{i,t}$	0.045***	0.045***	0.045***	0.045***	0.044***
	(22.043)	(21.470)	(21.541)	(21.529)	(21.487)
$\Delta Squared\_Return_{i,t-1}$	0.023***	0.023***	0.023***	0.023***	0.023***
	(15.745)	(15.465)	(15.368)	(15.293)	(15.242)
$\Delta Spread_{i,t}$	-0.058***	-0.058***	-0.058***	-0.057***	-0.057***
	(-10.178)	(-10.116)	(-10.027)	(-9.908)	(-9.918)
$\Delta Spread_{i,t-1}$	-0.056***	-0.056***	-0.056***	-0.055***	-0.055***
	(-15.534)	(-15.394)	(-15.226)	(-15.157)	(-15.174)
$\Delta GSV_{i,t}$		0.002	0.002	-0.003	-0.007
		(0.177)	(0.158)	(-0.212)	(-0.480)
$\Delta GSV_{i,t-1}$			-0.003	-0.002	-0.007
			(-0.237)	(-0.189)	(-0.631)
$\Delta TV_{i,t}$				0.041***	0.052***
				(4.955)	(5.863)
$\Delta TV_{i,t-1}$					0.033***
					(4.105)
Constant	-0.500***	0.333***	-0.497***	0.340***	-0.492***
	(-20.458)	(9.615)	(-20.013)	(9.772)	(-20.077)
Time Fixed Effects	YES	YES	YES	YES	YES
Company Fixed Effects	YES	YES	YES	YES	YES
R-squared	0.633	0.633	0.633	0.634	0.635
N	10,297	10,068	10,008	9,995	9,993

Column (1) to (5) report the result on market entry: $\Delta Volatility_{i,t} = \alpha + \beta_1 \Delta Volatility_{i,t-1} + \beta_2 \Delta \text{SquaredReturn}_{i,t} + \beta_3 \Delta \text{SquaredReturn}_{i,t-1} + \beta_4 \Delta \text{Turnover}_{i,t-1} + \beta_5 \Delta \text{Turnover}_{i,t-1} + \beta_6 \Delta \text{Spread}_{it} + \beta_7 \Delta \text{Spread}_{i,t-1} + \beta_8 \Delta GSV_{i,t} + \beta_9 \Delta GSV_{i,t-1} + \beta_{10} \Delta TV_{i,t} + \beta_{11} \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}$. Column (1) includes the control variables changes in lagged turnover, changes in squared returns, changes in lagged squared returns, as well as changes in spreads and lagged spreads. In Column (2), the variable GSV is added. Column (3) adds the lagged change in GSV. Column (4) adds the change in TV. Column (5) adds the lagged changes in TV. The dependent variable ΔVolatility stands for the percentage changes in volatility. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. All regressions are panel fixed effects regression, including time and company fixed effects. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10%, 5% or 1% levels, respectively.

3.5.3 High and low trading activity

The information GSV and TV incorporate have a higher impact if changes in trading activity are high. We sort the different values for daily trading volume of the 29 DJIA stocks in five groups. These groups range from low trading volume (0%-20%) to high trading volume (80%-100%). For each group we perform our standard regression on changes in turnover (7). In Table 3.9, we report the results for the

top 20% and lowest 20% of observations according to trading volume. Our results show that the impact of changes in GSV and TV on changes in turnover is higher (lower) for high (low) trading volume. This indicates that investor sentiment and investor attention are higher for high trading volume. The impact of TV is higher for the top 20% (0.246) than for the lowest 20% (0.062) but both times significant at the 1% level. The impact of changes in GSV and lagged GSV is only significant for the high volume stocks.

We assume that more informed traders trade low volume stocks, as there is a smaller or no impact of GSV and TV. For high volume stocks we assume to observe less informed traders but more noise traders. This is in-line with Easley et al. (1996) who find that the probability of informed trading is higher (lower) among low (high) volume stocks.

Table 3.9 Market entry – high and low trading volume

	(1) 20% lowest volume	(2) 20% highest volume
VARIABLES	$\Delta Turnover_t$	$\Delta Turnover_t$
$\Delta Turnover_{i,t-1}$	-0.368***	-0.362***
	(-25.040)	(-21.535)
$\Delta Squared\_Return_{i,t}$	0.019***	0.035***
	(9.672)	(15.954)
$\Delta Squared\_Return_{i,t-1}$	0.011***	0.018***
	(6.34)	(9.378)
$\Delta GSV_{i,t}$	0.009	0.097***
	(0.66)	(5.45)
$\Delta GSV_{i,t-1}$	0.001	0.044**
	(0.067)	(2.618)
$\Delta TV_{i,t}$	0.062***	0.246***
	(6.134)	(17.594)
$\Delta TV_{i,t-1}$	0.042***	0.134***
	(4.999)	(10.819)
Constant	0.096***	0.357***
	(5.775)	(10.632)
Time Fixed Effects	YES	YES
Company Fixed Effects	YES	YES
R-squared	0.66	0.641
N	4,736	5,066

Column (1) and (2) report the result on market entry: $\Delta Turnover_{i,t} = \alpha + \beta_1 \Delta Turnover_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta GSV_{i,t} + \beta_5 \Delta GSV_{i,t-1} + \beta_6 \Delta TV_{i,t} + \beta_7 \Delta TV_{i,t-1} + \mu_i + \lambda_t + v_{i,t}$. Column (1) shows the lowest 20% and column (2) the highest 20% in terms of trading volume. All regressions are panel fixed effects regression, including time and company fixed effects. The dependent variable ΔTurnover stands for the percentage changes in trading activity. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. We use Driscoll and Kraay standard errors.*, **, *** indicate statistical significance at the 10%, 5% or 1% levels, respectively.

3.5.4 Industry and sector group effects

We classify the different stocks into industry and 9 sector groups according to the profile from Yahoo finance.[41] The sectors are Industrials (17%), Technology (17%), Consumer Cyclicals (14%), Financial Services (14%), Consumer Defensive (10%), Healthcare (10%), Communication services (7%), Energy (7%) and Basic Materials (4%).

Running our standard panel fixed effects regression on changes in turnover (3.7) and volatility (3.8) with time, company fixed effects and sector fixed effects or time and sector fixed effects, does not lead to significant results. Here, the sectors have no impact. Running separately our regressions for all nine sectors, the results are comparable to the results we obtain in Tables 3.3 and 3.4[42]. As we have only 29 companies in the panel, the company fixed effects and the industry fixed effects are very similar or even coincide.

For changes in GSV and TV, the results are different. First, the results do not turn out significant for all sectors. Second, for some sectors the sign of the coefficient changes. However, the results are better for changes in TV than for changes in GSV. One explanation is that the media coverage of the sectors can differ on Google and Twitter.

3.5.5 Arellano Bond measure

In equation (3.7) the dependent variable is turnover in equation (3.8) it is volatility. As control variable we include the lagged dependent variable on the right hand side. The impact of the lagged dependent variable on the dependent variable is captured through β_1. The coefficient $\beta_1$turns out significant at the 1% level in both equations and captures the impact of the lagged dependent variable on the dependent variable.

The problem is that the lagged dependent variable can lead to first order autocorrelation and endogeneity problems (Roodman, 2006). We estimate our standard regressions on changes in turnover (3.7) and volatility (3.8) applying the approach suggested by Arellano and Bond (Arellano & Bond, 1991;

[41] https://finance.yahoo.com/quote/A/profile?p=A (30.8.2018, example Apple). We only look at the effects of the nine different sectors as there are 24 industry groups.
[42] The results on sector level are reported in the Appendix 3 and 4.

Roodman, 2006). To eliminate individual effects we use first differences of the dependent variable. Further, we restrict the set of independent control variables to the first differences of the lag of the dependent variable, changes in GSV and TV at time t. Our results are presented in Table 3.10, they show that our results are persistent.

In column (1), we present the results of regression (3.7) on changes in turnover. We find that the lagged changes in turnover have a negative and significant impact on changes in turnover (-0.1581%). The impact of changes in GSV (TV) on the dependent variable remain positive and significant at the 5% (1%) level. To confirm the validity of the instruments, we apply to overconfidence tests, the Sargan test and the Hansen test (Roodman, 2006). Both tests confirm the validity of the instruments.

In column (2), we present the results of regression (3.8). We find that an increase in yesterday's volatility leads to a decrease of volatility today. Moreover, we find that the impact of GSV on changes in volatility is weaker (0.097) than the impact of TV (0.246) but both remain highly significant.

Table 3.10 Market entry and noise traders – Arellano Bond

VARIABLES	(1) $\Delta\text{Turnover}_t$	(2) $\Delta\text{Volatility}_t$
$\Delta\text{Turnover}_{t-1}$	-0.1581***	
	(-7.34)	
$\Delta\text{Volatility}_{t-1}$		-0.2884***
		(-15.66)
ΔGSV_t	0.8247**	0.097***
	(2.19)	(5.45)
ΔTV_t	0.7160***	0.246***
	(7.53)	(17.594)
Constant	-0.0006**	-0.0008
	(-2.09)	(-0.83)
Arellano Bond test for AR(1)	0.0000	0.000
Arellano Bond test for AR(2)	0.002	0.049
Sargan test	0.016	0.096
Hansen test	0.250	0.271
N	25,494	25,494

Column (1) reports the results for the equation on changes in turnover $\Delta Turnover_{i,t} - \Delta Turnover_{i,t-1} = \alpha + \beta_1(\Delta Turnover_{i,t-1} - \Delta Turnover_{i,t-2}) + \beta_2(\Delta GSV_{i,t} - \Delta GSV_{i,t-1}) + \beta_3(\Delta TV_{i,t} - \Delta TV_{i,t-1}) + (v_{it} - v_{it-1})$. Column (2) reports the results for the equation on changes volatility $\Delta Volatility_{i,t} - \Delta Volatility_{i,t-1} = \alpha + \beta_1(\Delta Volatility_{i,t-1} - \Delta Volatility_{i,t-2}) + \beta_2(\Delta GSV_{i,t} - \Delta GSV_{i,t-1}) + \beta_3(\Delta TV_{i,t} - \Delta TV_{i,t-1}) + (v_{it} - v_{it-1})$. The dependent variable ΔTurnover stands for the percentage changes in trading activity. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10%, 5% or 1% levels, respectively.

3.5.6 Fama-French factors

To see if the effect we find for changes in GSV and TV is persistent, we control include the three standard Fama-French factors for the US market (Fama & French, 1993): Excess return (RMRF), stock performance of small stocks compared to big stocks (SMB) and the performance of value stocks compared to growth stocks (HML).

The correlation between the Fama-French Factors and the other control variables is reported in Table 3.11. The correlation of the three Fama-French Factors with the other variables is always below 30 %. The condition number using scaled variables lies at 2.09.

Table 3.11 Descriptive statistics including Fama-French Factors

	$\Delta Turnover_t$	$\Delta Volatility_t$	$\Delta Squared\_Return_t$	$Rm\text{-}Rf_t$	SMB_t	HML_t
ΔTurnover	1.0000					
ΔVolatility	0.4967***	1.0000				
ΔSquared_Return	0.3301***	0.4358***	1.0000			
RMRF	-0.1296***	-0.1177***	-0.0107	1.0000		
SMB	-0.0098	-0.0387***	-0.0100	0.1885***	1.0000	
HML	0.0466***	0.0440***	0.0172***	-0.0317***	-0.2268***	1.0000

Fama-French factors: Market excess return (RMRF), Small minus Big (SMB) and High minus Low (HML) . ΔGSV and ΔTV stand for percentage changes in GSV and TV. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. ΔVolatility stands for the percentage changes in volatility. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

In Table 3.12 we report the results for regression (3.7) on changes in turnover including the Fama-French factors. All three Fama-French factors are significant at the 1 % level in columns (1) to (5). As expected from the correlation analysis, we find a negative impact of the market excess return (RMRF) and the Small minus Big (SMB) on changes in turnover. For the book-to-market factor (HML), we find a positive impact on changes in turnover. The impact of changes in GSV (TV) on changes in turnover remains significant at the 1 % level.

Table 3.12 Market entry – including Fama-French factors

VARIABLES	(1) $\Delta Turnover_t$	(2) $\Delta Turnover_t$	(3) $\Delta Turnover_t$	(4) $\Delta Turnover_t$	(5) $\Delta Turnover_t$
$\Delta Turnover_{i,t-1}$	-0.296***	-0.294***	-0.299***	-0.287***	-0.341***
	(-36.870)	(-36.226)	(-38.018)	(-35.472)	(-50.642)
$\Delta Squared\_Return_{i,t}$	0.038***	0.038***	0.038***	0.033***	0.031***
	(30.040)	(30.514)	(30.715)	(32.702)	(32.835)
$\Delta Squared\_Return_{i,t-1}$	0.020***	0.020***	0.019***	0.018***	0.016***
	(22.971)	(22.927)	(22.795)	(23.751)	(22.907)
$RMRF_{i,t}$	-0.175***	-0.134***	-0.167***	-0.172***	-0.182***
	(-219.544)	(-147.454)	(-115.615)	(-153.792)	(-189.426)
$SMB_{i,t}$	-0.205***	-0.145***	-0.194***	-0.183***	-0.175***
	(-166.366)	(-125.973)	(-97.903)	(-95.435)	(-84.588)
$HML_{i,t}$	0.024***	-0.212***	0.025***	0.034***	0.046***
	(107.126)	(-87.059)	(82.915)	(69.253)	(46.405)
$\Delta GSV_{i,t}$		0.071***	0.095***	0.066***	0.051***
		(8.470)	(9.156)	(8.017)	(6.633)
$\Delta GSV_{i,t-1}$			0.049***	0.045***	0.027***
			(6.463)	(6.322)	(4.073)
$\Delta TV_{i,t}$				0.147***	0.173***
				(23.164)	(24.357)
$\Delta TV_{i,t-1}$					0.092***
					(18.787)
Constant	0.176***	0.025***	0.171***	0.165***	0.149***
	(112.063)	(29.048)	(89.821)	(96.033)	(68.377)
Time Fixed Effects	YES	YES	YES	YES	YES
Company Fixed Effects	NO	NO	NO	NO	NO
R-squared	0.492	0.495	0.497	0.545	0.560
N	25,566	24,970	24,833	24,811	24,801

Column (1) to (5) report the result on market entry: $\Delta Turnover_{i,t} = \alpha + \beta_1 \Delta Turnover_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 RmRf_{i,t} + \beta_5 SMB_{i,t} + \beta_6 HML_{i,t} + \beta_7 \Delta GSV_{i,t} + \beta_8 \Delta GSV_{i,t-1} + \beta_9 \Delta TV_{i,t} + \beta_{10} \Delta TV_{i,t-1} + \lambda_t + v_{i,t}$. Column (1) includes the control variables changes in lagged turnover, changes in squared returns, changes in lagged squared returns and the three Fama-French factors[43]. Market excess return (RMRF), Small minus Big (SMB) and High minus Low (HML). In column (2), the variable GSV is added. Column (3) adds the lagged change in GSV. Column (4) adds the change in TV. Column (5) adds the lagged change in TV. All regressions are Panel fixed effects regression, including time and company fixed effects. All regressions are panel fixed effects regression, including time and company fixed effects. The dependent variable ΔTurnover stands for the percentage changes in trading activity. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

The impact of the Fama-French factors on changes in volatility are reported in table 3.13. The impact of the Fama-French factors on changes in volatility varies if we consider changes GSV and TV as control variables. The coefficients of RMRF, SMB and HML turn out as expected from the correlation analysis in column (1). As soon as we integrate changes in GSV in column (2), the impact of the market capitalization (SMB) becomes positive (0.310). If we integrate the lagged changes in GSV in column (4), the

[43] Source: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

coefficient becomes insignificant. The impact of the value and growth stocks (HML) is positive and significant in columns (1) and (2). If we add the lagged changes in GSV in column (3), the coefficient is no longer significant. Adding the changes in TV in column (4) turns the coefficient negative and significant at the 1 % level. It becomes insignificant as soon as we add the lagged changes in TV in column (5). The coefficient of changes in GSV are insignificant. Changes in TV are highly significant at the 1 % level.

Table 3.13 Share of noise trader – including Fama-French factors

VARIABLES	(1) $Volatility_t$	(2) $Volatility_t$	(3) $Volatility_t$	(4) $Volatility_t$	(5) $Volatility_t$
$\Delta Volatility_{i,t-1}$	-0.474***	-0.474***	-0.474***	-0.473***	-0.475***
	(-84.220)	(-82.158)	(-82.102)	(-82.266)	(-82.720)
$\Delta Turnover_{i,t}$	0.571***	0.574***	0.574***	0.550***	0.543***
	(44.498)	(44.177)	(44.195)	(41.828)	(40.159)
$\Delta Turnover_{i,t-1}$	0.250***	0.250***	0.250***	0.245***	0.232***
	(26.305)	(25.711)	(25.431)	(25.254)	(21.978)
$\Delta Squared\_Return_{i,t}$	0.047***	0.047***	0.047***	0.047***	0.047***
	(35.847)	(35.257)	(35.613)	(35.416)	(35.276)
$\Delta Squared\_Return_{i,t-1}$	0.022***	0.023***	0.022***	0.022***	0.022***
	(22.537)	(22.030)	(22.036)	(22.175)	(22.011)
$RMRF_{i,t}$	-0.113***	-0.046***	-0.098***	-0.101***	-0.105***
	(-47.749)	(-42.873)	(-50.663)	(-52.209)	(-44.345)
$SMB_{i,t}$	-0.020***	0.310***	0.001	0.004	0.003
	(-6.971)	(89.910)	(0.598)	(1.596)	(1.149)
$HML_{i,t}$	0.095***	0.104***	0.001	-0.015***	-0.007
	(134.385)	(52.073)	(0.195)	(-3.631)	(-1.479)
$\Delta GSV_{i,t}$		-0.004	-0.002	-0.007	-0.010
		(-0.613)	(-0.238)	(-0.857)	(-1.208)
$\Delta GSV_{i,t-1}$			0.004	0.004	0.001
			(0.555)	(0.533)	(0.077)
$\Delta TV_{i,t}$				0.036***	0.043***
				(7.558)	(8.386)
$\Delta TV_{i,t-1}$					0.020***
					(3.721)
Constant	-0.229***	-0.024***	-0.288***	-0.297***	-0.296***
	(-70.659)	(-4.526)	(-136.925)	(-127.088)	(-123.593)
Time Fixed Effects	YES	YES	YES	YES	YES
Company Fixed Effects	NO	NO	NO	NO	NO
R-squared	0.593	0.593	0.594	0.595	0.596
N	25,566	24,970	24,833	24,811	24,801

Column (1) to (5) report the result on market entry: $\Delta Volatility_{i,t} = \alpha + \beta_1 \Delta Volatility_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 RmRf_{i,t} + \beta_5 SMB_{i,t} + \beta_6 HML_{i,t} + \beta_7 \Delta GSV_{i,t} + \beta_8 \Delta GSV_{i,t-1} + \beta_9 \Delta TV_{i,t} + \beta_{10} \Delta TV_{i,t-1} + \lambda_t + v_{i,t}$. Column (1) includes the control variables changes in lagged turnover, changes in squared returns, changes in lagged squared returns and the three Fama-French factors[44]. Market excess return (RMRF), Small minus Big (SMB) and High minus Low (HML). In column (2) the variable GSV is added. Column (3) adds the lagged change in GSV. Column (4) adds the change in TV. Column (5) adds the lagged change in TV. The dependent variable ΔVolatility

[44] Source: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

stands for the percentage changes in volatility. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. All regressions are Panel fixed effects regression, including time and company fixed effects. The regressions are estimated with Driscoll and Kraay standard errors.*, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

3.5.7 Autocorrelation

It is possible that we have autocorrelation between changes in turnover (volatility) and lagged changes in turnover (volatility). If we include fixed effects, this could lead to biased estimation results of the error term and the other coefficients. We run the regressions (3.7) and (3.8) with and without time fixed effects and company fixed effects. We report the results of changes in turnover in Table 3.14 and the results on changes in volatility in Table 3.15.

In Table 3.14 we find that the constant changes from 0.000 in column (1) and (2) to -0,023 in column (3) and (4) when we integrate time fixed effects. Further, the size of the standard errors in absolute terms increases with time fixed effects. The R^2 of the regression increases from 0.29 to 0.56 as soon as we integrate time fixed effects. The size of the coefficients and the significance level are not affected.

Table 3.14 Market entry – without fixed effects

VARIABLES	(1) $\Delta Turnover_t$	(2) $\Delta Turnover_t$	(3) $\Delta Turnover_t$	(4) $\Delta Turnover_t$
$\Delta Turnover_{i,t-1}$	-0.355***	-0.355***	-0.341***	-0.341***
	(-26.742)	(-26.742)	(-50.642)	(-50.660)
$\Delta Squared\_Return_{i,t}$	0.039***	0.039***	0.031***	0.031***
	(28.888)	(28.890)	(32.835)	(32.837)
$\Delta Squared\_Return_{i,t-1}$	0.021***	0.021***	0.016***	0.016***
	(16.211)	(16.213)	(22.907)	(22.916)
$\Delta GSV_{i,t}$	0.092***	0.091***	0.051***	0.051***
	(7.532)	(7.529)	(6.633)	(6.627)
$\Delta GSV_{i,t-1}$	0.048***	0.048***	0.027***	0.027***
	(5.552)	(5.553)	(4.073)	(4.068)
$\Delta TV_{i,t}$	0.163***	0.163***	0.173***	0.173***
	(19.595)	(19.591)	(24.357)	(24.356)
$\Delta TV_{i,t-1}$	0.086***	0.086***	0.092***	0.092***
	(12.252)	(12.251)	(18.787)	(18.783)
Constant	-0.000	-0.000	0.023***	0.023***
	(-0.028)	(-0.028)	(6.403)	(6.232)
Time Fixed Effects	NO	NO	YES	YES
Company Fixed Effects	NO	YES	NO	YES
R-squared	0.2910	0.2912	0.5600	0.5605
N	24,801	24,801	24,801	24,801

Column (1) to (5) report the result on market entry: $Turnover_{it} = \alpha + \beta_1 Turnover_{it-1} + \beta_2 SquaredReturn_{it} + \beta_3 SquaredReturn_{it-1} + \beta_4 GSV_{it} + \beta_5 GSV_{it-1} + \beta_6 TV_{it} + \beta_7 TV_{i,t-1} + \varepsilon$. Column (1) does not include fixed effects.

Column (2) includes company fixed effects. Column (3) includes only time fixed effects. Column (4) includes time and company fixed effects. All regressions are panel regressions. The dependent variable ΔTurnover stands for the percentage changes in trading activity. . The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA.The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

In table 3.15 we find that the constant becomes negative and significant and R^2 increases to 0.56 when we include time fixed effects. The coefficients of the control variables change only slightly when we add time fixed effects in column (3) and (4). Their significance level stays at 1 %. The coefficients of changes in GSV and lagged changes in GSV are always insignificant. What we find is that the impact of TV and lagged TV on changes in volatility is higher and more significant if we include time fixed effects.

Table 3.15 Share of noise trader – without fixed effects

VARIABLES	(1) $Volatility_t$	(2) $Volatility_t$	(3) $Volatility_t$	(4) $Volatility_t$
$\Delta Volatility_{i,t-1}$	-0.457***	-0.457***	-0.475***	-0.475***
	(-52.039)	(-52.041)	(-82.720)	(-82.730)
$\Delta Turnover_{i,t}$	0.525***	0.525***	0.543***	0.543***
	(23.771)	(23.765)	(40.159)	(40.157)
$\Delta Turnover_{i,t-1}$	0.240***	0.240***	0.232***	0.232***
	(15.562)	(15.556)	(21.978)	(21.979)
$\Delta Squared\_Return_{i,t}$	0.048***	0.048***	0.047***	0.047***
	(25.738)	(25.743)	(35.276)	(35.283)
$\Delta Squared\_Return_{i,t-1}$	0.021***	0.021***	0.022***	0.022***
	(13.873)	(13.874)	(22.011)	(22.020)
$\Delta GSV_{i,t}$	0.009	0.009	-0.010	-0.010
	(0.831)	(0.825)	(-1.208)	(-1.213)
$\Delta GSV_{i,t-1}$	0.018	0.018	0.001	0.001
	(1.954)	(1.947)	(0.077)	(0.069)
$\Delta TV_{i,t}$	0.040***	0.040***	0.043***	0.043***
	(5.755)	(5.755)	(8.386)	(8.386)
$\Delta TV_{i,t-1}$	0.016*	0.016*	0.020***	0.020***
	(2.034)	(2.033)	(3.721)	(3.720)
Constant	0.000	0.000	-0.149***	-0.024***
	(0.068)	(0.068)	(-34.995)	(-4.053)
Time Fixed Effects	NO	NO	YES	YES
Company Fixed Effects	NO	YES	NO	YES
R-squared	0.4740	0.4744	0.5960	0.5957
N	24,801	24,801	24,801	24,801

Column (1) to (5) report the result on the share of noise traders: $\Delta Volatility_{i,t} = \alpha + \beta_1 \Delta Volatility_{i,t-1} + \beta_2 \Delta SquaredReturn_{i,t} + \beta_3 \Delta SquaredReturn_{i,t-1} + \beta_4 \Delta Turnover_{i,t} + \beta_5 \Delta Turnover_{i,t-1} + \beta_6 \Delta GSV_{i,t} + \beta_7 \Delta GSV_{i,t-1} + \beta_8 \Delta TV_{i,t} + \beta_9 \Delta TV_{i,t-1} + \lambda_t + v_{i,t}$. Column (1) does not include fixed effects. Column (2) includes company fixed effects. Column (3) includes only time fixed effects. Column (4) includes time and company fixed effects. All regressions are panel regressions. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

3.6 Risk of online investor sentiment

The rise of GSV and TV as a source of data has also weaknesses which cannot be neglected. We have to question if the information we obtain is not manipulated somehow.

Generally speaking, data manipulation is not a new phenomenon. Leinweber and Madhavan (2001) describe the history of 300 years of market manipulation. Successful manipulation depends on access to media, anonymity, scalability, time and impact of the technical opportunities developed. The technical opportunities these days bring the manipulation problem to a new level. Before the internet, only few people access to information and not everyone could publish information. Rumours spread through newspapers and special finance magazines. Nowadays, the internet can connect everyone which increases the audience. Different trading strategies, such as pump-and-dump or bluffing, attract especially uninformed retail investors while the informed investors leave the game at the peak with high profits. Whereby pump and dump applies in particular to penny stocks, the most recent example being game stop.

The Securities and Exchange Commission (SEC) has even an own office dealing with this kind of market manipulation and insider trading. They inform investors in their regular alerts on social media about fraud and stock rumours (SEC, 2014, 2015).

It is not clear to what extend Google and Twitter data is affected by manipulation. In 2016 the most popular social media scams were dating and romance. The loss due to scams is increasing (Commission, 2017). According to Gerth (2017), market manipulation with Twitter often miss the mark. There are more than one million stock relevant tweets on Twitter per day and manipulation is easy as a fake account, managed by humans or bots, is sufficient to commit fraud. Especially short selling attacks are a problem. This kind of market manipulation can affect the whole market. Data on these manipulations are hard to find. One example is the tweet by Elon Musk after which short seller lost approximately 1.3 billion USD (Bain & Mott, 2018). A recent report by the Federal Financial Supervisory Authority (Gillert, 2017) discussed the topic of short selling attacks or too positive mailings. They pointed out that wrong information no matter if positive or negative can lead to reputational damages. Retail investors have to verify information to protect themselves.

We can conclude that at the moment mail and phone are more affected than social media (Commission, 2017). Hence, we do not have to adjust for Twitter or Google manipulation in our data. But individual investors need to be aware as well as the regulation environment for the years to come.

3.7 Conclusion

In this chapter we analyse the effect of GSV and TV on trading activity and the share of noise traders on the market. We find that both contain new information and have predictive power on financial market indicators. We use a panel data setting of 29 DJIA stocks over a period of three years. Our results are robust to a variety of robustness checks and control variables.

First, we find that an increase in GSV and TV have a highly significant impact on trading activity. An increase in GSV (TV) by 10 % leads to an increase in turnover by 0.51 % (1.73 %). Further, we find that GSV and TV have predictive power. An increase in t-1 leads to an increase in turnover by of GSV (TV) by 0.27 % (0.92 %). Robustness test show that this effect is even more pronounced for high volume stocks in line with Easley et al. (Easley et al., 1996).

Second, we find that TV has a highly significant impact on volatility. An increase in TV by 10 % leads to an increase in volatility by 0.4 %. An increase in TV in t-1 leads to an increase in volatility in t by 0.2 % which shows the predictive power. In line with the DSSW model (De Long) we see the impact of TV on volatility as an indicator for an increase in the share of noise traders on the market. GSV has no impact on volatility.

Our work supports the idea that GSV and TV capture the beliefs of individual investors. In line with the findings of Mao et al. (2015) we find that TV has a higher impact on turnover and volatility than GSV. We assume that this lies in the nature of Twitter and Google. While Twitter is a microblogging platform for news and opinions, GSV only measures what people search for if it raised their attention. Therefore it makes sense to distinguish between investor attention for GSV and investor sentiment for TV. Further research is needed to confirm these finding based on additional internet data sources and different measures for investor beliefs.

The impact of Twitter and Google on Volatility - A Time-Series Analysis

Carolin Hartmann[45]

[45] University of Hohenheim, Institute of Financial Management, Mailing address: Schwerzstraße 38, 70599 Stuttgart, GERMANY, Tel: +49 (0) 7 11 45 92 29 00, Email: burghof@uni-hohenheim.de

4 The impact of Twitter and Google on Volatility - A Time-Series Analysis

4.1 Introduction

Google and Twitter have users worldwide and collect insights on user attention and sentiment data. Therefore, they can be considered as potential indicators for market movements. This chapter looks at and compares the impact of Google and Twitter on RV of the DJIA on a daily level for a three year period.

The basic idea is based on the behavioural approach that investors do not make their decisions in a purely rational way, as the efficient market hypothesis describes (Fama, 1970). Instead, other factors influence the decision of investors and affect asset pricing and volatility on financial markets. Kahneman and Tversky (Kahneman, 1973; Kahneman & Tversky, 1979; Tversky & Kahneman, 1974) lay the foundation in the early 1970s and show that the judgment of individuals is subject to errors which makes it difficult to measure behaviour.

A number of authors develop the idea of human behaviour in their studies. De Long et al. (1990) lay the theoretical foundation in an agent based model of rational traders and irrational "noise" traders- DSSW. They find that noise traders influence all investors and thus price formation on capital markets and volatility. Other authors build on this approach and confirm the influence of individual investors on capital markets (Alfarno & Lux, 2007; Barber, Odean, & Zhu, 2009; Hirshleifer, 2001; Hirshleifer & Teoh, 2003; Kumar, 2007; Kumar & Lee, 2006; Lux & Marchesi, 1999; Tetlock, 2007; Verma & Verma, 2007). Shleifer and Vishny (1997) add the limits to arbitrage, as institutional investors only act as a corrective to a certain extent. Baker and Wurgeler (2007) take up these two approaches and make investor sentiment, the influence of irrational investors, measurable. They take investor sentiment as an exogenous effect which is difficult to measure. Their study shows that investor sentiment has the most effect on stocks that are hard to value or difficult to arbitrage.

Baber and Odean (2008) find that individual investors prefer attention-grabbing stocks. Attention is achieved, for example, through media coverage or exceptionally high returns. However, institutional investors are not prone to attention, since they have more sophisticated ways to evaluate stocks.

Already in 2004 Antweiler and Frank analysed posted messages on the DJIA and the Dow Jones Internet Index, finding a positive correlation between message volume with trading volume and volatility. Wang et al. (2006) find in their study that market based sentiment does not improve volatility prediction. Sentiment measures become useless if returns are integrated in the forecasting.

Google and Twitter create new opportunities to track user behaviour and provide new measures of investor sentiment and attention. In the academic literature, it is widely discussed whether and how this publicly available information can improve volatility forecasts.

Da et al. (2011) are one of the first to use GSV to approximate human behaviour. Before, it was only possible to track publications, e.g. in newspapers and the media (see e.g. Barber, Brad M; Odean, 2008) but it was not clear what individuals pay attention to. Observing Google searches allows Da et al. (Da et al., 2011) to track the behaviour of individuals more closely. An increase in GSV means that individuals pay more attention to a certain topic. This finding is also applicable to financial market topics. The authors use GSV as a new and direct measure of investor attention. They find that an increase in GSV leads to higher returns in the next two weeks.

Further authors find a positive effect of GSV on volatility forecasts (Andrei & Hasler, 2015; Dimpfl & Jank, 2011, 2016; Hamid & Heiden, 2015; Vlastakis & Markellos, 2012; Vozlyublennaia, 2014).

Dimpfl and Jank (2011, 2016)[46] assess the impact of investor attention on volatility. They use Google Search queries on the DJIA to measure investor attention. To measure day-to-day RV they apply the measure presented by Andersen et al. (2003) with 10-minute return intervals. They find that Google search queries improve the prediction of volatility in-sample and out-of-sample. They find that during times of strong market movements the attention to stocks increases. A high level of searches today leads to an increase in volatility tomorrow. Applying a VAR model they show with the Granger causality test that searches cause volatility. These results hold especially in high volatility times.

[46] The first version in 2011 was a working paper which was published in a modified version in 2016

Vlastakis and Markellos (2012) use Google search queries on 30 companies listed on the NYSE and NASDAQ as an information proxy. They find a positive relationship with historical and implied volatility and trading volume. Especially in times of high returns, the demand for information is higher.

Vozlyublennaia (2014) provides evidence on the relationship between past stock returns and attention of stock market indices such as DJIA, NASDAQ index and the S&P 500 with volatility. She shows that investor attention granger cause returns and provides evidence on the short term effect of attention on returns. For volatility the results are less persistent, there is an impact of volatility on attention but not vice versa. Compared to other studies (see Barber, Brad M; Odean, 2008; Da et al., 2011) Vozlyublennaia (2014) finds positive and negative price pressure.

Hamid und Heiden (2015) find that on a weekly level their model outperforms in-sample and out-of-sample forecasts of traditional autoregressive models. These results do not hold on a daily level. Especially in phases of high volatility investor attention improves forecasts in the short-run. As Dimpfl and Jank (2011) the authors use the realized variance of the Dow Jones index based on 5-minute return intervals.

Andrei and Hasler (2015) proxy investor attention to financial and economic news with GSV. They find a positive relationship between investor attention with volatility and risk premium. High investor attention leads to a faster in-pricing of information and a high risk premium compared to low attention stocks.

Besides Google, authors also use the microblogging platform Twitter to analyse investor behaviour. In addition to the number of tweets, tweets can also be classified according to the mood they reflect or simply as positive and negative sentiment.

Zhang et al. (2011) use tweets to predict stock market indicators. Their finding: Emotional tweets containing the words “hope”, “fear” and “worry” reflect uncertainty. These emotional tweets can be predictors of the stock market. They are negatively correlated with stock market performance the next day, but positively correlated with volatility[47].

[47] Chicago Board Options Exchange volatility index on the S&P 500 (VIX).

Bollen et al. (2011) study if public mood expressed on twitter can influence and predict capital market indicators. On a daily basis they create a twitter mood time-series on the DJIA with various mood types. One of their findings is that some mood types granger cause closing prices of the DJIA three to four days later. Further, forecasting accuracy increases with the use of a trained fuzzy neural network.

According to Sprenger et al. (2014) tweets are potential proxies for investor sentiment. Their analysis shows a correlation between sentiment and financial market indicators such as return and volatility. Further, an increase in bullishness is an indicator for an increase in stock prices. Their interpretation is that the information of microblogs has not fully reached the market.

Mao et al. (2015) look at both, Google queries and Twitter messages. They create a bullishness index reflecting the frequency of the terms “bullish” and “bearish” on both platforms for large stock market indices such as the DJIA. The authors use weekly data from Google and daily Twitter data. One of their findings is that Twitter bullishness has a statistical and economic effect on stock market indices. This indicates that Twitter captures new, not priced information. Also the Google bullishness index is correlated with financial indicators but has no statistically significant predictive power. Their findings indicate a lead-lag relationship between Twitter bullishness and Google bullishness, which underpins Twitter an advantage over Google. In the sense of De Long et al. (1990) Twitter bullishness is a possible investor sentiment proxy for noise trader. Even if they find correlation, causality is difficult to identify and to maintain. The authors see the need to increase the measurement accuracy for Google and Twitter data.

Oliveira et al. (Oliveira, Cortez, & Areal, 2013) use data on five large US stocks and the S&P 500 from the microblogging Stock Twits platform to see if they can improve forecasts on returns, volatility and trading volume. They only find an effect of posting volume on trading volume forecasts. Their sentiment variables do not improve forecasting of the other financial market indicators. In a later study Oliveira et al. (Oliveira, Cortez, & Areal, 2017) use aggregated daily data from different microblogging platforms and create attention and sentiment indicators for US stock market indices. Even if they find that certain sentiment indicators improve volatility forecasts, the model is not more accurate that the autoregressive baseline model.

Behrendt and Schmidt (2018) analyse the effect of Twitter sentiment and activity on intraday stock return volatility on the DJIA. They find that stock-related tweets do not improve out-of-sample volatility forecasts. What they find is a co-movement of tweets and the DJIA intraday volatility. In a more recent paper Ballinari and Behrendt (2020) point out that investor sentiment from Twitter and StockTwits explain the problem through undetected structural breaks in the data.

Audrino et al. (2020) use GSV and Stock Twits message to create sentiment and attention variables. As suggested in earlier research they combine different indicators to improve forecasting (e.g. Mao et al., 2015; Oliveira et al., 2017; Sprenger et al., 2014). Both variables have predictive power on future volatility, even if the improvements are economically small. They find that this is especially the case for companies with a high market capitalization. Good indicators are the number of search engine queries or the posting volume on social media.

4.2 Methodology

From a research perspective it is not finally clarified if data from Google and Twitter contain valuable information on the behaviour of investors which is not yet incorporated in financial markets. Results if Google and Twitter data improve volatility forecasting in the literature are mixed, a correlation is probable. Only certain aspects of Google and Twitter seem to impact volatility. Furthermore, the time horizon and the index in question play an important role. Google and Twitter allow to capture the beliefs of individual investors. But there is still no valid and persistent measure of Google and Twitter data. The research idea of this chapter is to compare Google and Twitter mostly following the idea of Dimpfl and Jank (Dimpfl & Jank, 2011), Mao et al. (Bollen et al., 2011; Hamid & Heiden, 2015; Mao et al., 2015) based on the general idea of De Long et al. (1990).

This chapter provides an in-depth time series analysis of GSV and TV on the RV of the DJIA over a period of three years from January 2014 until December 2016. The hypothesis is that GSV and TV contain new information to improve volatility forecasts. While microblogs such as Twitter stand for objectivity, Google searches benefit from an enormous user base (Sprenger et al., 2014).

The hypothesis is tested in a two-step procedure. First, I show that GSV and TV improve the in-sample model. Based on these results the second step shows that the GSV data and the TV improve out-of-sample forecasts. For robustness checks various control variables are included in the model. Significance of the Google and Twitter variables emphasize the effect that valuable information is contained in the Google and Twitter data. As a result I can show that the volume of GSV and TV contain publicly available information that optimize forecasts.

4.3 Data and descriptive statistics

4.3.1 Dataset and sample

The data covers a time period of three years, from 5 January 2014 up to 30 December 2016. The database is established from five different sources. First, GSV on the DJIA is extracted from Google Trends[48]. Second, TV on the DJIA is from Sowa Labs[49]. Third, Intraday data on the DJIA is from Thomson Reuters Tick History[50]. Fourth, Datastream is the source for daily data on the DJIA, data on the volatility index of the DJIA (VXD) and data on DJIA Futures. Fifth, Data on the three month US Treasury Bills is from the FED of the Louisiana. The analysis is conducted on a daily level.

RV pictures the market movement of the DJIA. It is calculated on a daily level based on intraday returns following the approach of Andersen et al. (Andersen et al., 2003):

$$RV_t = \sqrt{\sum_{j=1}^{n} r_{t,j}^2}, \tag{4.1}$$

with $r_{t,j}^2$ standing for the squared returns at day t for the interval j. The number of intraday return intervals is measured by n. To calculate RV, I use ten minutes intervals between 9:30 am up to 4 pm which leads to $n = 39$ observations per day. The length of the interval is optimal to avoid microstructure effects (see e.g. Andersen et al., 2003; Dimpfl & Jank, 2011). To obtain a well behaved RV time series Andersen et al. (2003) suggest to take the logarithmic form of RV. Moreover following Andersen et al.

[48] https://trends.google.de/trends/?geo=DE (last access 10.06.2019)
[49] https://www.sowalabs.com/
[50] The data was downloaded and evaluated in 2016.

(2003) week-ends and holidays are excluded from the sample to avoid the modelling of specific week-end effects. Other studies use similar approaches (Areal & Taylor, 2002; Audrino et al., 2020; Dimpfl & Jank, 2016; Hamid & Heiden, 2015; Wang et al., 2006).

Overall there are 68,492 return observations in the sample. The intraday returns range from -0.0249973 up to 0.0206534 with an overall mean close to the expected mean return of zero (0.00000543). The Dickey-Fuller test confirms the stationarity of the intraday returns (Dickey & Fuller, 1979). There are no intraday trends. As in the literature I use log RV in the regressions (Dimpfl & Jank, 2011; Wang et al., 2006).

GSV data on the DJIA is from Google Trend. The procedure is comparable to other papers e.g. Da et al. (2011), Dimpfl and Jank (2011) Fink and Johann (2014), Hamid and Heiden (2015). GSV is a time series which depicts the relative search volume of a specific search term on Google during a certain time span or point in time. In this chapter I use daily data. To specify the search and get the least disturbance as possible, I apply the search term "Dow Jones Industrial Average Index". With the topic market index I also filter for search terms related to the DJIA. Further, I use the filter "finance" to narrow the search on financial topics. The search volume index is scaled to the maximum GSV of the whole period, which is the 9 November 2016. This day marks the election of the 45th American president Donald Trump. Further, I look at the 30 stocks of the DJIA over time and establish a second GSV index with the search criteria on the 30 stocks. In the end two variables are obtained from the Google database, a variable on the DJIA index and a variable summing up all 30 stocks of the DJIA during the period of interest.

TV is a variable from Sowa Labs[51], a former FinTech start-up with trained algorithms to evaluate Twitter Tweets. The algorithm filters with respect to keywords and hashtags on the DJIA. The data on the DJIA is available from the 1 January 2014 up to the 31 December 2016. Peter et al. (2017) describe and use a comparable dataset for a different timespan. Following their results, I use the total amount of TV on the DJIA without separating the tweets in positive, negative or neutral sentiment data. All three variables are depicted in figure 4.1 below.

[51] https://www.sowalabs.com/

Figure 4.1 Realized volatility, Google search volume and Twitter volume

RV of the DJIA (left) GSV (middle) and TV (right) on the DJIA from June 2014 to January 2017.

To capture market expectations in the forecasting I control for financial and macroeconomic factors. First, because earlier research pointed out that they have an impact on future volatility. Second, because I want to distil the marginal impact of Google and Twitter on volatility in in-sample and out-of-sample predictions. I use four control variables: DJIA Futures, implied volatility of the DJIA (VXD), interest rate of three month US treasury bills and daily DJIA returns.

Various studies consider trading volume for their volatility prediction (Dimpfl & Jank, 2011; Oliveira et al., 2013, 2017; Vlastakis & Markellos, 2012; Wang et al., 2006), but it is calculated historically for a specific trading day. I consider DJIA futures to capture market expectations to forecast volatility. In the research literature, futures are a source of information for the underlying spot markets which can lead to an increase in volatility (Antoniou & Holmes, 1995). They can improve the prediction of returns (Brooks et al., 2001). Further, stock index futures also have a stabilizing effect on stock markets which can increase market efficiencies (Bologna & Cavallo, 2002). The future variable I integrate in the regression is the E-mini future. A future contract which is a portion of standard DJIA futures[52]. Datastream offers a continuous and daily time series on the traded DJIA volume called "CBT-Mini Dow Jones Continuous"[53] [54].

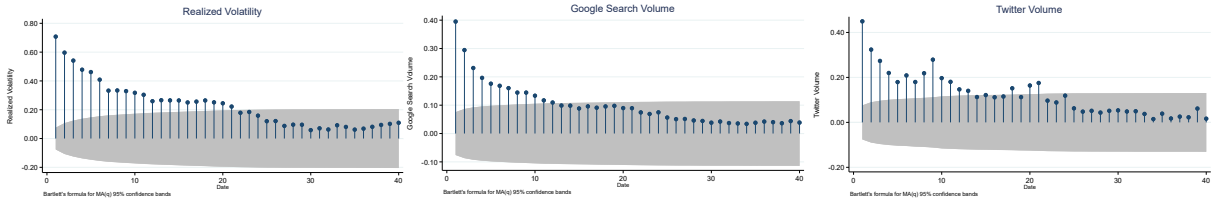
To control for volatility I use the implied volatility of the DJIA (VXD)[55]. It is calculated by the Chicago Board Options Exchange on options of the DJIA on a daily basis. As the DJIA future it is forward looking and captures expectations about the development of the DJIA volatility. To capture the current risk free interest rate I use the 3-Month Treasury Bill (T3) from the federal reserve of Louisiana[56]. To

[52] They are traded at the Chicago Mercantile Exchange's Globex electronic trading platform and the New York Board of Trade
[53] CYMCS0.
[54] https://www.cnbc.com/quotes/?symbol=@DJ.1 https://www.cmegroup.com/trading/equity-index/us-index/e-mini-dow_quotes_globex.html BNP
[55] www.cboe.com/VXD 3.12.2018
[56] https://fred.stlouisfed.org/series/DTB3

control daily returns of the DJIA I calculate the percentage change per day based on the DJIA daily closing prices data from Datastream. Earlier research pointed out that the inclusion of return data make the forecasting with sentiment variables obsolete (Antweiler & Frank, 2004; Wang et al., 2006)

Figure 4.2 depicts the autocorrelation of the variables RV, GSV and TV. For all three it shows a decaying dependency. While the autocorrelations is significant in the beginning, it decreases over time becoming insignificant. The largest change is visible between t and t+1.

Figure 4.2 Autocorrelation of Realized volatility, Google search volume and Twitter volume



Autocorrelation of DJIA (left graph), GSV (middle graph) and TV (right graph) on the DJIA. The grey area marks 95 % confidence interval.

Due to multicollinearity it is not possible to include all four control variables, GSV and TV in the same regression.[57] To solve this issue I apply a sequential Ordinary Least Square (OLS) regression. Combining the four macroeconomic variables step-by-step I obtain the residual $R3$, a control variable which combines the characteristics of all four macroeconomic variables.

To measure the impact of GSV (TV) I continue the sequential regression and obtain a residual for GSV (TV). For GSV (TV) I obtain the residual Gr (Tr). It captures the effect of GSV (TV) which remains after controlling for the four macroeconomic variables. As a last step, I create a residual variable $G2r$ ($T2r$) to see the effect of GSV (TV) if I control for TV (GSV) and the macroeconomic factors. By construction the condition number of the residuals is one, there is no multicollinearity left.

4.3.2 Summary statistics

Table 4.1 Panel A reports the descriptive statistics of the different residuals ($R3$, Gr, Tr, $G2r$, $T2r$) and RV. The correlation between the macroeconomic residual $R3$ and RV is negative and significant but not

[57] Appendix 5 shows the descriptive statistics on the original variables.

high (-11.67 %). As $R3$ is a variable of combined residuals, it is not possible to disentangle the effect and to say which variable of the four control variables is the major driver of the result. The correlation between residuals of GSV (Gr and $G2r$) and TV (Tr and $T2r$) are high but not problematic as the variables are not part of the same regression.

In Panel B of Table 4.1 the summary statistics show the behaviour of the different variables. All residuals have a mean of zero by construction. The behaviour of the macroeconomic residual (R3) is close to normal. The residuals of GSV and TV are positively skewed and have a high kurtosis. The data is not corrected to keep all possible effects in the data. In the literature it is pointed out that especially in high volatility times Google and Twitter data can make a difference (Dimpfl & Jank, 2011; e.g. Hamid & Heiden, 2015), therefore especially the deviations are of interest.

Table 4.1 Descriptive statistics

Panel A: Correlation Analysis						
	log_RV	R3	Gr	Tr	T2r	G2r
RV	1.0000					
R3	-0.1167***	1.0000				
Gr	0.0479	0.0000	1.0000			
Tr	0.0282	-0.0000	0.5152***	1.0000		
T2r	0.0041	-0.0000	0.0000	0.8571***	1.0000	
G2r	0.0390	0.0000	0.8571***	-0.0000	-0.5152***	1.0000

Panel B: Summary Statistics							
	Mean	S.D.	Skewness	Kurtosis	Min.	Max.	N
RV	-5.37	0.46	0.30	3.18	-6.76	-3.48	668
R3	0.00	0.15	0.76	2.39	-0.19	0.40	667
Gr	0.00	4.07	17.41	378.29	-6.91	90.92	667
G2r	0.00	3.49	14.34	299.56	-11.14	73.51	667
Tr	0.00	47.31	4.72	45.98	-92.19	575.87	667
T2r	0.00	40.55	3.69	34.03	-151.64	462.53	667

Panel A: Correlation of log RV (RV), control variable ($R3$), a residual of the macroeconomic and financial factor variables, residuals of GSV ($Gr, G2r$), residuals of TV ($Tr, T2r$). Panel B summary statistics. All variables are calculated on a daily basis from 5 January 2014 until 30 December 2016 for DJIA. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

4.4 Empirical study

4.4.1 Basic model

To test the impact of GSV and TV on RV, I run two different models. The findings show that GSV (Gr_{t-1}) and TV (Tr_{t-1} and $T2r_{t-1}$) have a significant impact on volatility (RV_t).

The model is a standard vector auto regressive (VAR) model by Sims (1980). Due to the hybrid form between a univariate time series model and a simultaneous equation model, I can analyse the impact of GSV (TV) on RV and vice versa. Further, in contrast to autoregressive models, RV can depend on its own lags but also on lags of GSV (Gr, $G2r$) and TV (Tr, $T2r$). In the regression I integrate GSV and TV measures but I can disentangle the effect to see if attention measured by GSV or sentiment measured by TV has a higher impact on RV. As exogenous variable I integrate the macroeconomic variable ($R3$). Equation (4.1) shows the standard VAR model with an exogenous variable. The dependent variable is y_t it depends on a constant α, its own lag y_{t-1} two lagged independent variables x_{1t-1} and x_{2t-1}, an exogenous variable x_{3t} and an error term u_t:

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 x_{1t-1} + \beta_3 x_{2t-1} + \beta_4 x_{3t} + u_t \quad (4.2)$$

$$RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t \quad (4.3)$$

$$RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Tr_{t-1} + \beta_3 G2r_{t-1} + \beta_4 R3_t + u_t. \quad (4.4)$$

Model 1 in equation (4.2) looks at the impact of RV_{t-1}, Gr_{t-1}, $T2r_{t-1}$ on RV_t, with a constant β_1, the exogenous variable ($R3_t$) and the error term u_t. Model 2 in equation (4.3) looks at the impact of RV_{t-1}, Tr_{t-1}, $G2r_{t-1}$ on RV_t, with a constant α, the error term u_t and the exogenous variable of residuals ($R3_t$).

The optimal lag length is determined by comparing different information criteria and is set to one.[58] This is in line with the literature and reflects the result of the autocorrelation in figure 4.2, which indicates that the major impact is in the first lag. In the appendix 6 and 7 different lag lengths are reported as well.

Table 4.2 reports the results for model 1 and 2. For both, the impact of the lagged dependent variable (RV_{t-1}) on RV_t is highly significant at the 1 % level (0.708). As expected from Figure 4.1, there is a high dependency of RV on its lags.

For model 1 granger causality tests support the result, Gr_{t-1} and $T2r_{t-1}$ influence RV_t but not the other way round. There is a negative impact of GSV (Gr_{t-1}) on RV today (RV_t), the coefficient β_2 in model

[58] Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (SBIC), and the Hannan and Quinn information criterion (HQIC)

1 is negative and significant at the 5 % level (-0.0083). The impact of TV ($T2r_{t-1}$) is still significant at the 10 % level and also negative (-0.00077).

For model 2 the results are comparable. Granger causality shows that Tr_{t-1} causes RV_t but not the other way round, $G2r_{t-1}$ does not cause RV_t. The impact of TV (Tr_{t-1}) on RV_t is negative and significant at the 1 % level (-0.00093), whereas there is no impact of GSV ($G2r_{t-1}$).

Overall, the residuals of GSV (Gr_{t-1}) and TV (Tr_{t-1}, $T2r_{t-1}$) effect RV, even if the control variable ($R3$) and the lagged dependent variable are in the equation. The basic model indicates that there is predictive power of GSV and TV on RV. This support the hypothesis that GSV and TV have an impact on volatility. They incorporate new, valuable information for the market.

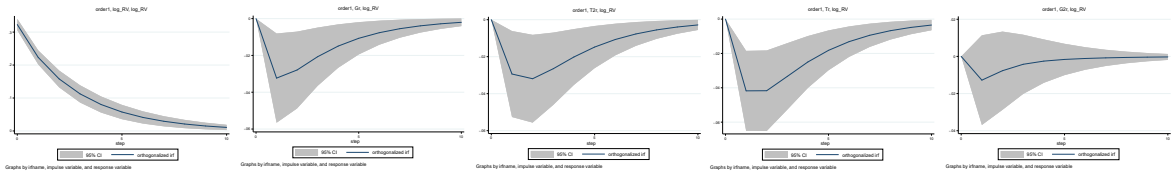
Table 4.2 VAR basic model

	Model 1			Model 2		
	(1)	(2)	(3)	(1)	(2)	(3)
	RV_t	Gr_t	$T2r_t$	RV_t	Tr_t	$G2r_t$
RV_{t-1}	0.70809***	-0.22300	-1.97497	0.70809***	-3.31031	-0.07626
	(25.93434)	(-0.65651)	(-0.60716)	(25.93434)	(-0.87131)	(-0.26221)
Gr_{t-1}	-0.00830**	0.15183***	0.10146			
	(-2.69890)	(3.96615)	(0.27675)			
$T2r_{t-1}$	-0.00077*	0.00576	0.31607***			
	(-2.48117)	(1.50160)	(8.59912)			
Tr_{t-1}				-0.00093***	0.30232***	-0.00244
				(-3.51751)	(8.21357)	(-0.86498)
$G2r_{t-1}$				-0.00372	-1.08865*	0.16558***
				(-1.03679)	(-2.17985)	(4.33097)
R3	-0.09589	-0.12708	-0.71926	-0.09589	-1.48023	-0.06147
	(-1.12333)	(-0.11967)	(-0.07073)	(-1.12333)	(-0.12462)	(-0.06760)
Constant	-1.56757***	-1.19758	-10.59905	-1.56757***	-17.77021	-0.40986
	(-10.65652)	(-0.65440)	(-0.60480)	(-10.65652)	(-0.86815)	(-0.26157)
N	666	666	666	666	666	666
RMSE	0.32	4.03	38.60	0.32	45.09	3.45
LL	-192	-1870	-3380	-192	-3480	-1770
R^2	0.51	0.03	0.10	0.51	0.10	0.03
AIC	16.36	16.36	16.36	16.36	16.36	16.36

For model 1 and 2 the table reports the results of a basic VAR regression. For model 1: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$ for model 2: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Column (1) reports results for log RV (RV). Column (2) reports for model 1 (model 2) the result for the residuals of GSV (TV) Gr_t (Tr_t)). Column (3) reports for model 1 (model 2) the result for the residuals of TV (GSV) $T2r_t$ ($G2r_t$)).As exogenous variable all regressions include R3, a combined residual of the macroeconomic and financial factor variables. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL). *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

The impulse response functions (IRF) in figure 4.3 visualize the effect of a shock in the endogenous variables (RV_{t-1}, Gr_{t-1}, $G2r_{t-1}$, Tr_{t-1} and $T2r_{t-1}$) on RV_t.[59] The five graphs show that all variables have an impact on RV_t, what differs are sign and magnitude of the shocks. A shock in RV_{t-1} leads to an increase in RV_t which declines over time. Comparable to the autocorrelation depicted in figure 4.2. The effect of shocks in Gr_{t-1} and Tr_{t-1} is much smaller and negative. Both lead to a decrease in RV_t but the effect is more pronounced for TV. Same holds for $G2r_{t-1}$ and $T2r_{t-1}$ but their impact is smaller both than for Gr_{t-1} and Tr_{t-1}. The results are in line with the results reported in table 4.2.

Figure 4.3 Impulse response functions



Impulse response functions of Model 1. Impact of RV_{t-1} on RV_t (left), impact of Gr_{t-1} on RV_t (2nd from left), impact of $T2r_{t-1}$ on RV_t (middle). Model 2, impact of Tr_{t-1} on RV_t (2nd from right), impact of $G2r_{t-1}$ on RV_t (right). The grey area indicates the 95 % confidence interval.

4.4.2 In-sample forecasting

As a next step I conduct a two-step procedure. First, an in-sample test, excluding a holdout set. Second, a forecasting approach on the holdout set. Results show that GSV and TV improve the goodness of fit in-sample and also improve out-of-sample forecast in extreme volatility periods.

The in-sample period covers one third of the entire period (222 days). Three different models test the impact of GSV and TV on RV. First I run the standard regressions for model 1 and model 2 and as benchmark, a standard AR (1) process as model 3. Equation (4.4) describes model 1 including a second lag for the independent variables RV_{t-2}, Gr_{t-2} and $T2r_{t-2}$. Equation (4.5) stands for model 2 extended by a second lag of the independent variables RV_{t-2}, Tr_{t-2} and $G2r_{t-2}$. Equation (4.6) describes a simple autoregressive process (AR (1)) with the dependent variable X_t, a constant β_0, its own lag X_{t-1} and an error term ε_t being identically and independently distributed (white noise). Equation (4.7) shows the AR (1) benchmark model.

[59] see Lütkepohl (2005, 51–63) and Hamilton (1994, 318–323)

$$RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 RV_{t-2} + \beta_3 Gr_{t-1} + \beta_4 Gr_{t-2} + \beta_5 T2r_{t-1} + \beta_6 T2r_{t-2} + \beta_7 R3_t + u_t \quad (4.5)$$

$$RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 RV_{t-2} + \beta_3 Tr_{t-1} + \beta_4 Tr_{t-2} + \beta_5 G2r_{t-1} + \beta_6 G2r_{t-2} + \beta_7 R3_t + u_t \quad (4.6)$$

$$X_t = \alpha + \beta_1 X_{t-1} + \varepsilon_t \quad (4.7)$$

$$RV_t = \alpha + \beta_1 RV_{t-1} + \varepsilon_t \quad (4.8)$$

Second, based on regressions (4), (5) and (7), I predict the in-sample $\widehat{RV_t}$ for all three models with a linear prediction and compare the results with RV_t. The results are depicted in table 4.3. Model 1 and 2 have similar results in terms of goodness of fit and slightly outperform the AR (1) process depicted in model 3. Compared to the AR (1) model they perform better, as they incorporate more information. The root mean squared error (RMSE) for all three models is close to zero, but the models with GSV and TV have slightly smaller RMSE terms (0.0014) than the AR (1) process (0.0017) in model 3. Looking at the goodness of fit measure (R^2) model 1 and 2 explain 63 % of the variation of RV while the AR (1) process stays lower with 48 %. The log-likelihood value indicates that as well, with a slightly higher value for model 1 and 2 than for model 3.

Table 4.3 In-sample VAR

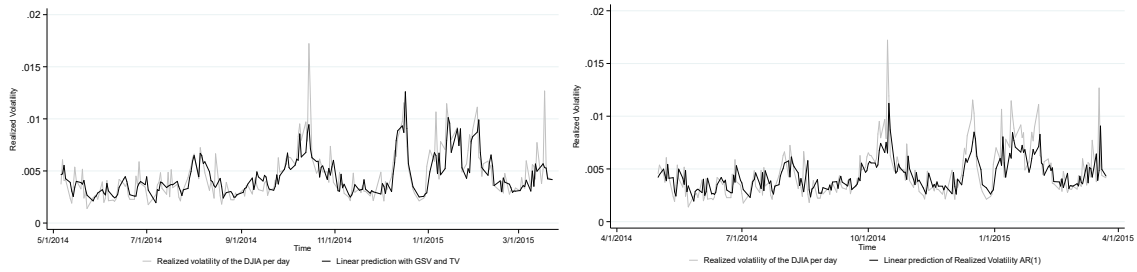
	In-sample (222 days)		
	RMSE	R^2	LL
Model 1	0.0014	0.6367	1120
Model 2	0.0014	0.6367	1120
Model 3	0.0017	0.4845	1100

Comparison of the in-sample linear predictions of predicted RV ($\widehat{RV_t}$) with actual RV (RV_t) for model 1, 2 and 3 with 222 days in-sample. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV ($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. Model 3 is a standard AR (1) process $X_t = \alpha + \beta_1 X_{t-1} + \varepsilon_t$. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL).

The graphical representation of the in-sample results in figure 4.4 shows the difference between model 1 and model 3.[60] In both graphs, the in-sample linear prediction follows the movement of the DJIA RV. But model 1 which includes GSV and TV captures better the extreme values than the AR (1) process.

[60] Model 2 leads to the same results than model 1 and is therefore not depicted. It is important that GSV and TV are integrated but it makes no difference in the forecasting in which combination this is done.

Figure 4.4 In-sample VAR



In-sample predictions with 222 days Model 1 (left), model 3 (right).

The academic literature indicates that sentiment and attention measures are especially valuable in volatile time (Dimpfl & Jank, 2011; Hamid & Heiden, 2015; see e.g. Vlastakis & Markellos, 2012). To measure the effect Dimpfl and Jank (2011, 2016) use the goodness of fit measures and depict them for different times of volatility. They find that in high volatility times GSV improves volatility predictions(Dimpfl & Jank, 2011).

Following the approach of Dimpfl and Jank (2011) I sort the RV values from high to low volatility and build nine volatility groups from the highest 1 % down to the lowest 5 %. The goodness of fit measures in each group (RMSE, R^2 and Log Likelihood) compare the predicted $\widehat{RV_t}$ values from models 1, 2 and 3 with the observed RV values (RV_t). Results are reported in table 4.4.

Looking at the top and bottom 50 % with respect to the goodness of fit measure R^2 supports the idea that GSV and TV have an impact on RV, especially in more volatile times. The impact of GSV and TV on volatility is higher in the top 50 % with an R^2 of 47 % for model 1 and 2 and 31 % for model 3. For the bottom 50 % R^2 is at 27 % for model 1 and 2 and at 12 % for model 3.

For the top 1 % of RV the impact of GSV and TV is highest with an R^2 of 70 %. The result of the AR (1) process (model 3) is much lower with only 15 %. GSV and TV improve the model fit in times of high volatility. For the top 5 % (10 %) the goodness of fit (R^2) decreases for all three models below 3 %. For the top 25 % (50 %) an R^2 of 27 % (48 %) for model 1 and 2, compared to only 16 % (31 %) for model 3.

Looking at the bottom, the effect of GSV and TV is less pronounced with an R^2 always below 30 % for all three models. But still R^2 for model 1 and 2 is always above R^2 for model 3.

In times of high volatility the impact of GSV and TV and the incorporated information is higher than in less volatile times, where the difference between the models 1 and 2 and the AR(1) process is less pronounced and the goodness of fit in terms of R^2 decreases. All three models exhibit the same pattern only the magnitude of RMSE, R^2 and LL differ. Models 1 and 2 lead to the same results, the order in which GSV and TV are taken into account is not important.

Table 4.4 In-sample VAR – high to low volatility

	Model 1			Model 2			Model 3		
	RMSE	R^2	LL	RMSE	R^2	LL	RMSE	R^2	LL
Top 1	0.0022	0.6975	16	0.0022	0.6975	16	0.0037	0.1512	14
Top 5	0.0022	0.0048	53	0.0022	0.0048	53	0.0022	0.0101	58
Top 10	0.0022	0.0242	105	0.0022	0.0242	105	0.0022	0.0024	109
Top 25	0.0019	0.2692	267	0.0019	0.2692	267	0.0020	0.1556	268
Top 50	0.0016	0.4756	550	0.0016	0.4756	550	0.0019	0.3136	546
Bottom 50	0.0005	0.2703	674	0.0005	0.2703	674	0.0006	0.1216	674
Bottom 25	0.0003	0.1074	356	0.0003	0.1074	356	0.0003	0.0704	361
Bottom 10	0.0002	0.0841	148	0.0002	0.0841	148	0.0002	0.0115	154
Bottom 5	0.0003	0.0019	70	0.0003	0.0019	70	0.0002	0.0001	77

Comparison of the in-sample linear predictions of predicted RV ($\widehat{RV_t}$) with actual RV (RV_t) for model 1, 2 and 3 with 222 days in-sample. Sorted in nine volatility groups ranging from high (top 1%) to low (bottom 5%) volatility. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV ($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. Model 3 is a standard AR (1) process $X_t = \alpha + \beta_1 X_{t-1} + \varepsilon_t$. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL).

4.4.3 Out-of-sample forecasting

The second step is forecasting on the hold out set with 447 days. The model specifications stay the same as in the in-sample prediction. Model 1 is represented by equation (4.4), model 2 by equation (4.5) and model 3 remains the standard AR (1) process of equation (4.7). The forecasting for all three models is a linear one day ahead forecast. As expected, the goodness of fit (R^2) and the RMSE are lower (higher) than for the in-sample prediction for all three models. The results in table 4.5 show that the AR (1) process in model 3 outperforms models 1 and 2 which include information on GSV and TV.

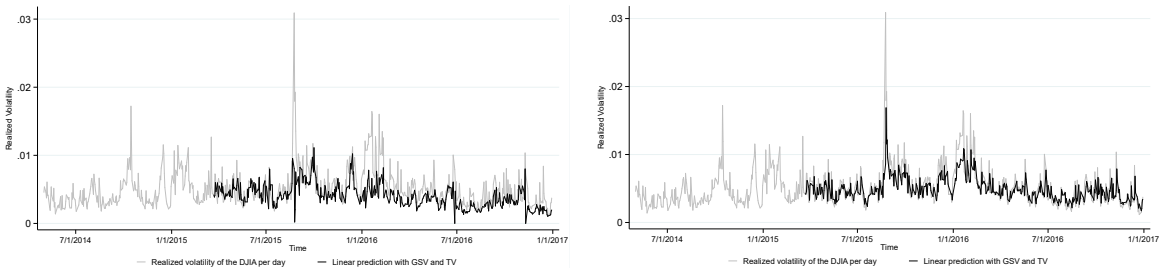
Table 4.5 Out-of-sample VAR

	Out-of-sample (447 days)		
	RMSE	R^2	LL
Model (1)	0.00248	0.28356	2040
Model (2)	0.00248	0.28356	2040
Model (3)	0.00209	0.48974	2120

Comparison of the out-of-sample linear predictions of predicted RV ($\widehat{RV_t}$) with actual RV (RV_t) for model 1, 2 and 3 with 447 days out-of-sample. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV ($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. Model 3 is a standard AR (1) process $X_t = \alpha + \beta_1 X_{t-1} + \varepsilon_t$. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit is measures R^2 and the log likelihood (LL) measure.

Figure 4.4 depicts the out-of-sample forecast. While the RV values of the AR (1) process in model 3 follows the true RV values closely, model 1 points in the wrong direction for some values. The AR(1) process of model 3 captures the movements of RV but the forecasted values are less pronounced than the true values of RV. Model 1 also follows the RV movement, but some peaks of the predicted RV values point in the opposite direction. The results show that there is an impact of GSV and TV but it is not precise enough to forecast RV. One explanation is that the measurement of sentiment and attention is not precise enough. I cover only the volume of GSV and TV and not the sentiment direction of the TV. This might explain why the peaks are detected but for certain values not the right direction in the prediction.

Figure 4.5 Out-of-sample VAR



Linear Out-of-sample predictions with 447 days out-of-sample. Model 1 (left), AR (1) process of model 3 (right)

Table 4.6 shows the results for ten different volatility groups starting from the top 1 % of RV down to the bottom 5 % of RV[61]. As in table 4.4, I compare for each group the out-of-sample predictions RV^ of model 1, 2 and 3 with the true RV value. Overall, the goodness of fit measures are lower than for the in-sample forecasts. For very high RV values and low RV values, GSV and TV improve out-of-sample forecasts. For less extreme values the AR (1) process performs better.

The goodness of fit measures indicate that for the top 50 %, the AR (1) process of model 3 outperforms model 1 and 2. This is in line with the observations from figure 4.4 which shows that especially the

[61] Due to more observations it is possible to build 10 groups instead of nine as for the in-sample prediction.

upside peaks are not always met by model 1and 2. Only for the top 1 % of RV, model 1 and 2 improve the goodness of fit with an R^2 of 15 % and the RMSE of 0.0066 compared to model 3. Like in the in-sample case this observation is persistent also in the out-of-sample forecasting. For all other ranges of top realized volatilities, model 3 outperforms model 1 and 2. For the bottom 50 % of RV the results differ, here model 1 and 2 fit as good as model 3 or better. For the lowest 1 % model 1 (2) outperform the AR (1) process.

Table 4.6 Out-of-sample VAR – high to low volatility

	Model 1			Model 2			Model 3		
	RMSE	R^2	LL	RMSE	R^2	LL	RMSE	R^2	LL
Top 1	0.0066	0.1533	19	0.0066	0.1533	19	0.0071	0.0132	19
Top 5	0.0044	0.0165	93	0.0044	0.0165	93	0.0042	0.0947	94
Top 10	0.0038	0.0091	188	0.0038	0.0091	188	0.0034	0.1985	193
Top 25	0.0032	0.0337	484	0.0032	0.0337	484	0.0028	0.2714	500
Top 50	0.0028	0.0979	998	0.0028	0.0979	998	0.0023	0.3905	1041
Bottom 50	0.0007	0.2705	1308	0.0007	0.2705	1308	0.0007	0.3000	1313
Bottom 25	0.0005	0.1658	697	0.0005	0.1658	697	0.0005	0.1277	694
Bottom 10	0.0004	0.1365	288	0.0004	0.1365	288	0.0004	0.1000	287
Bottom 5	0.0003	0.2322	148	0.0003	0.2322	148	0.0003	0.1633	147
Bottom 1	0.0001	0.6417	31	0.0001	0.6417	31	0.0002	0.0114	29

Comparison of the out-of-sample linear predictions of predicted RV ($\widehat{RV_t}$) with actual RV (RV_t) for model 1, 2 and 3 with 447 days out-of-sample. Sorted in ten volatility groups ranging from high (top 1%) to low (bottom 5%) volatility. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV ($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. Model 3 is a standard AR (1) process $X_t = \alpha + \beta_1 X_{t-1} + \varepsilon_t$. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL).

4.5 Robustness test

4.5.1 Individual effect of GSV and TV

In the basic model equations (4.2) and (4.3), I include the residuals for GSV and TV in both equations. In equation (4.17) and (4.18) I include only the first lag of the GSV residual (Gr_{t-1}) and the TV residual (Tr_{t-1}).

$$RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 R3_t + u_t \tag{4.9}$$

$$RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Tr_{t-1} + \beta_3 R3_t + u_t. \tag{4.10}$$

Table 4.7 shows that the VAR model results do not change compared to table 4.2, the result of the VAR model is consistent. I find that the impact of Gr_{t-1} (Tr_{t-1}) is negative, of the same size and significant at the 5 % (10 %) level as in the basic model. What changes is the constant in both models β_1 decreases. Further, the coefficient β_3 of the TV residual (-0.00093) is smaller than the coefficient β_3 of GSV residual (-0.00830) but it is highly significant at the 1 % level. The coefficient of GSV in equation (4.8) is only significant at the 5 % level. This supports the idea that GSV and TV have an impact on RV but it is rather small.

Table 4.7 VAR basic model – separate effect of GSV and TV

	Model 1		Model 2	
	(1) RV_t	(2) Gr_t	(1) RV_t	(2) Tr_t
RV_{t-1}	0.70780***	-0.22081	0.70695***	-3.6427
	-25.80494	(-0.64897)	-25.89272	(-0.95616)
Gr_{t-1}	-0.00830**	0.15183***		
	(-2.68638)	-3.95937		
Tr_{t-1}			-0.00093***	0.30241***
			(-3.51351)	-8.1869
R3	-0.09591	-0.12689	-0.09716	-1.85108
	(-1.11847)	(-0.11929)	(-1.13738)	(-0.15530)
Constant	-1.56912***	-1.18594	-1.57368***	-19.55695
	(-10.61816)	(-0.64695)	(-10.69800)	(-0.95281)
N	666	666	666	666
RMSE	0.3253	4.0350	0.3241	45.2157
LL	-195	-1870	-193	-3480
R^2	0.5075	0.0233	0.5112	0.0921
AIC	6.2210	6.2210	11.0440	11.0440

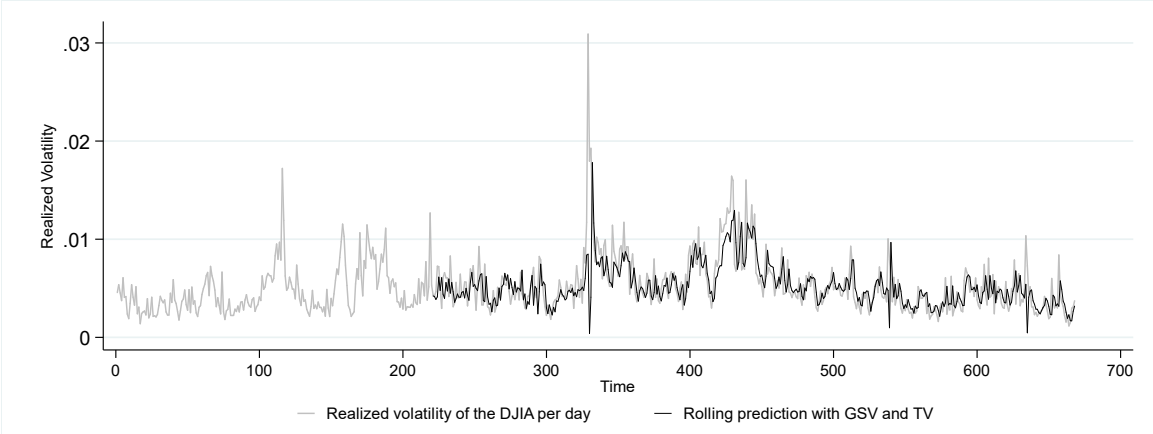
For model 1 and 2 the table reports the results of a basic VAR regression. For model 1: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$ for model 2: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Column (1) reports results for log RV (RV). Column (2) reports for model 1 (model 2) the result for the residuals of GSV (TV) Gr_t (Tr_t)). As exogenous variable all regressions include R3, a combined residual of the macroeconomic and financial factor variables. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL). *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

4.5.2 Rolling window out-of-sample forecast

Changing the forecasting method from linear prediction to rolling window prediction increases the goodness of fit (R^2) for model 1 from 28 % to 37 % but this is still lower than the AR (1) process of model 3 with 49 %. Figure 4.5 shows the result of the rolling prediction with equation (4) integrating GSV and TV. The result is comparable with the linear prediction results of figure 4.4. The predicted RV values resemble more closely the true movement of the RV, but some peaks tend in the wrong directions as for

the linear predictions. This means that the VAR model with GSV and TV detects the extreme value but for certain peaks it is not correctly reflected in the prediction.

Figure 4.6 Out-of-sample VAR – rolling window prediction



Rolling out-of-sample predictions with 447 days out-of-sample. Model 1 (left), AR (1) process of model 3 (right)

Further, I compare the outcome in times of high and low volatility. I find that for the top 5 % of the rolling forecast outperformance the linear forecast in terms of goodness of fit (4.23 %). Same hold for the top 50 % here the R^2 of the Rolling forecast lies at 20.1 % while the linear forecast lies at 9.8 %. The RMSE and the Log Likelihood measures are in all cases comparable with the linear predictions in table 4.6. But still none of the rolling window predictions including GSV and TV outperform the simple AR (1) of model 3 in table 4.6.

Table 4.8 Out-of-sample VAR rolling window – high to low volatility

	Model 1			Model 2		
	RMSE	R^2	LL	RMSE	R^2	LL
Top 1	0.0071	0.0087	19	0.0071	0.0087	19
Top 5	0.0044	0.0423	93	0.0044	0.0423	93
Top 10	0.0038	0.0011	188	0.0038	0.0011	188
Top 25	0.0032	0.0760	491	0.0032	0.0760	491
Top 50	0.0027	0.2063	1016	0.0027	0.2063	1016
Bottom 50	0.0007	0.3027	1320	0.0007	0.3027	1320
Bottom 25	0.0005	0.1621	702	0.0005	0.1621	702
Bottom 10	0.0004	0.1422	289	0.0004	0.1422	289
Bottom 5	0.0003	0.2236	148	0.0003	0.2236	148
Bottom 1	0.0002	0.0657	29	0.0002	0.0657	29

Comparison of the out-of-sample rolling window predictions of predicted RV ($\widehat{RV}_t$) with actual RV (RV_t) for model 1 and 2 with 447 days out-of-sample. Sorted in ten volatility groups ranging from high (top 1%) to low (bottom 5%) volatility. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV

($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL).

Table 4.9 shows a comparison of the linear and rolling window prediction for model 1 and 2 following equations (4.4) and (4.5). I find that for both models the linear predictions lead to better results in terms of goodness of fit. The Root Mean Square Error (RMSE) for the linear (rolling) prediction is at 0.0024 (0.0589). Same holds for the Mean Absolut Error (MAE) and the mean absolute percent error. Both are also smaller for the linear prediction in model 1 and 2 compared to the rolling prediction. Also Theil's U indicates that the linear prediction performs better than a naïve[62] prediction, as it is below 1 for model 1 and 2 and above 1 for the rolling window approach.

Table 4.9 Out-of-sample VAR comparison – linear and rolling window prediction

	Model 1		Model 2	
	Linear	Rolling	Linear	Rolling
RMSE	0.0024	0.0589	0.0024	0.0589
MAE	0.0014	0.0046	0.0014	0.0046
MAPE	0.2637	0.9230	0.2637	0.9230
Theil's U	0.9046	19.6946	0.9046	19.6946

Comparison of the out-of-sample linear and rolling predictions of predicted RV ($\widehat{RV_t}$) with actual RV (RV_t) for model 1 and 2 with 447 days out-of-sample. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV ($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit is measures R^2 and the log likelihood (LL) measure.

4.5.3 Variation of the forecast period

The forecasting so far takes 222 days in-sample and 447 days out-of-sample. Keeping all other things equal I change to 334 days in-sample and 335 days out-of-sample. Model 1 follows equation (4.4), model 2 equation (4.5) and model 3 remains the AR (1) process described in equation (4.7). The RMSE (0.0019) and the Log Likelihood (1620) are now comparable for all three models. The goodness of fit measures (R^2) is even better for model 1 and 2 with 50.42 % than for model 3 with 50.34 %.

[62] Theil's U=1: Naïve prediction would mean to simple take the last value as forecast.

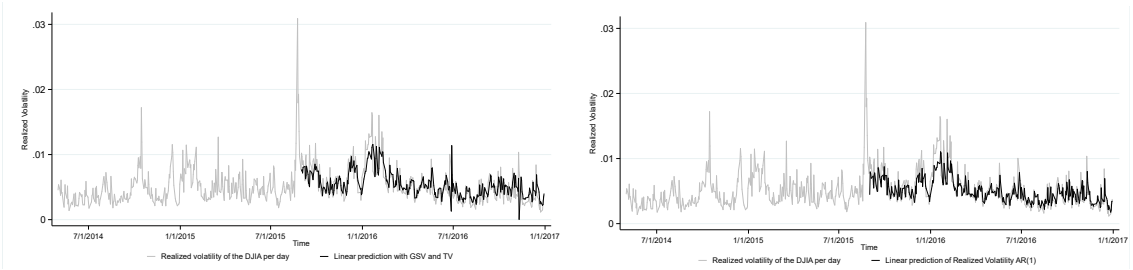
Table 4.10 Out-of-sample VAR – shorter time period

	Out-of-sample (335 days)		
	RMSE	R^2	LL
Model 1	0.0019	0.5042	1620
Model 2	0.0019	0.5042	1620
Model 3	0.0019	0.5034	1620

Comparison of the out-of-sample linear predictions of predicted RV ($\widehat{RV_t}$) with actual RV (RV_t) for model 1, 2 and 3 with 335 days out-of-sample. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV ($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. Model 3 is a standard AR (1) process $X_t = \alpha + \beta_1 X_{t-1} + \varepsilon_t$. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit is measures R^2 and the log likelihood (LL) measure.

The RV time-series is less volatile for the shorter out-of-sample forecasting period. Figure 4.7 shows that with the shorter out-of-sample time span, the wrongly measured peaks by model 1 and 2 decrease which can be an explanation for the better overall fit of both models compared to the AR (1) process.

Figure 4.7 Out-of-sample VAR – shorter time-period



Linear out-of-sample predictions with 335 days out-of-sample. Model 1 (left), AR (1) process of model 3 (right)

The shorter out-of-sample forecasting period also impacts the high and low volatility comparison. While the goodness of fit for the top 1 % is low for all three models, the top 5 %, top 10 % and top 25 % shows a better model fit of model 1 and 2 compared to model 3. In this constellation we see the positive impact of GSV and TV on high RV. For the top 50 % the three models are comparable with an R^2 of approximately 36 %. For the low volatility values the three models have comparable results with an R^2 of 30 % (bottom 50 %), decreasing to 14 % for the AR (1) process and 0.04 % for model 1 and 2. The impact of GSV and TV is lower for less volatile periods.

Table 4.11 Out-of-sample VAR shorter time period – high to low volatility

	Model 1			Model 2			Model 3		
	RMSE	R^2	LL	RMSE	R^2	LL	RMSE	R^2	LL
Top 1	0.0016	0.0314	21	0.0016	0.0314	21	0.0016	0.0435	21
Top 5	0.0017	0.1130	85	0.0017	0.1130	85	0.0018	0.0451	85
Top 10	0.0017	0.2467	170	0.0017	0.2467	170	0.0018	0.1365	167
Top 25	0.0020	0.2374	404	0.0020	0.2374	404	0.0021	0.1779	401
Top 50	0.0019	0.3565	807	0.0019	0.3565	807	0.0019	0.3693	813
Bottom 50	0.0007	0.3070	975	0.0007	0.3070	975	0.0007	0.3144	981
Bottom 25	0.0005	0.1412	519	0.0005	0.1412	519	0.0005	0.1297	518
Bottom 10	0.0004	0.1274	217	0.0004	0.1274	217	0.0004	0.1057	216
Bottom 5	0.0003	0.1967	107	0.0003	0.1967	107	0.0003	0.1280	106
Bottom 1	0.0003	0.0037	22	0.0003	0.0037	22	0.0003	0.1409	22

Comparison of the out-of-sample linear predictions of predicted RV ($\widehat{RV_t}$) with actual RV (RV_t) for model 1, 2 and 3 with 335 days out-of-sample. Sorted in ten volatility groups ranging from high (top 1%) to low (bottom 5%) volatility. Model 1 measures the impact of residuals for GSV (Gr_{t-1}) and TV ($T2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Model 2 measures the impact of the residuals for TV (Tr_{t-1}) and GSV ($G2r_{t-1}$) on RV, following the equation $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 Gr_{t-1} + \beta_3 T2r_{t-1} + \beta_4 R3_t + u_t$. Both models contain as exogenous variable R3, a combined residual of the macroeconomic and financial factor variables. Model 3 is a standard AR (1) process $X_t = \alpha + \beta_1 X_{t-1} + \varepsilon_t$. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL).

4.6 Conclusion

Researchers examine in the current academic literature whether and how investor sentiment and investor attention influence financial factors (see e.g. Audrino et al., 2020; Behrendt & Schmidt, 2018; Dimpfl & Jank, 2016; Hamid & Heiden, 2015). Combination of different internet search volumes and platforms have an influence on financial indicators such as return, volatility and trading volume. Especially in the current literature the measures for investor sentiment and investor attention are combined (see e.g. Audrino et al., 2020). The literature finds that GSV, TV and other internet sources are correlated with financial factors, but the extent to which GSV and TV have an economically significant impact is not yet conclusively assessed.

In this chapter I use GSV and TV to proxy investor attention and investor sentiment. I assess their predictive power on the RV (RV) of the DJIA in-sample and out-of-sample. To control for financial and macroeconomic factors, I use four variables: Interest rate, implied volatility of the DJIA, trading volume of DJIA futures and DJIA returns. This approach is important to ensure that the effect of GSV and TV contains new information and does not cover existing market information. To avoid multicollinearity I calculate residuals for the independent variables in a sequential OLS regression. I obtain residual of GSV ($Gr, G2r$), TV ($Tr, T2r$) and a combined variable of the four control factors ($R3$). While Gr (Tr)

measures the effect of GSV (TV) controlling for the four control variables. $G2r$ ($\mathrm{T}2r$) measures the effect of GSV (TV) remains, if I control for the control variables and TV (GSV). RV (RV) is in logarithmic form.

The classical VAR model with one lag shows that the residuals of GSV ($Gr, G2r$) and the residuals of TV ($Tr, T2r$) granger cause RV. There is no effect of RV on the residuals of GSV and TV. The impact of the residuals is relatively small but significant at the 1 % to 5 % level. The influence of TV is more important, as it is more significant. This result is consistent with the literature (see e.g. Mao et al., 2015). The small values of the coefficients lie in the work with residuals.

To see whether this relationship is also true for predictions, I perform an in-sample and then an out-of-sample estimation of a classical OLS regression model with a linear one day ahead forecast. For the in-sample predictions models including GSV and TV outperform a classic AR (1) process in terms of goodness of fit. Looking at high and low RV, the prediction with GSV and TV are especially better for the top 1 %. For the out-of-sample GSV and TV do not predict better than the AR (1) process in terms of goodness of fit. For different volatility classes, models including GSV and TV outperform the AR (1) process for the extreme values, the top 1 % and the bottom 1 %. Especially in extreme situations GSV and TV can incorporate new information. The robustness check with a shorter out-of-sample period improves the results. This underlines that the results of GSV and TV are not generally persistent but depend on the selected criteria. The robustness check with a rolling window approach instead of a linear one day ahead forecast do not improve the results.

Overall, I find an impact of GSV and TV on RV but the impact is not as persistent as it could be. Measures for investor attention and investor sentiment have room for improvement. Therefore, future research should have three points in mind. First, the question on how to measure GSV and TV and other platform and search volume data is important. The availability of data and its granularity is constantly improving and thus allows for further insights. With respect to Twitter, the granularity might help to improve the detection of specific mood types. Second, the use of different models using neural networks or machine learning approaches might detect connection that are not covered by standard time-series

models. Third, the combination of different measures of investor sentiment and investor attention might improve the precision and the persistency of measures for investor beliefs.

5 Conclusion and Outlook

5.1 Conclusion

This thesis shows that investor beliefs have an impact on financial market indicators. Investor beliefs have the potential to increase the understanding of financial markets and have an impact on the prediction of future market movements. The findings of the three chapters show that differences in implied volatility, GSV and TV are proxy variables for investor behaviour. Causality and economic significance must be proven in the long term.

In chapter 2, we were able to introduce a new variable to measure investor sentiment. It compares the implied volatility measures on the DAX at the Frankfurt Stock Exchange (VDAX and VDAX-NEW) and our implied volatility index (VSSE) for the Stuttgart Stock Exchange. The sentiment measure makes the difference between retail investor behaviour at the SSE and professional investors at the FSE visible. It is significant in predicting the daily returns on a size-based long-short portfolio. Our analysis shows the persistent inconsistence between prices of structured products for retail investors on the Stuttgart Stock Exchange (SSE) and option prices of professional investors on the Frankfurt Stock Exchange (FSE).

The results remain significant if we calculate different implied volatility measures for the SSE or use another measure for the implied volatility at the FSE. All these results provide empirical evidence that there are significant persistent behavioural differences between the two investor types. Our sentiment indicator partially captures this difference and the persistent mispricing.

In chapter 3, we find that GSV and TV contain new information and have predictive power on a daily level. The impact of TV on financial markets is more important than the impact of GSV. We use a daily panel data setting of 29 DJIA stocks over a period of three years. Our work supports the idea that GSV and TV capture the beliefs of individual investors. First, we show that changes in GSV and TV have an impact on trading activity. On the same day, an increase in GSV (TV) by 10 % leads to an increase in turnover by 0.51 % (1.73 %). An increase of GSV (TV) in t-1 leads to an increase in turnover at time t by 0.27 % (0.92 %). Second, we find that TV has a highly significant impact on RV. An increase in TV

in at time t by 10 % leads to an increase in volatility in t by 0.4 %. An increase in TV in t-1 leads to an increase in RV at time t by 0.2 % which shows the predictive power of TV. In line with the DSSW model (De Long et al., 1990) we see the impact of TV on volatility as an indicator for an increase in the share of noise traders on the market. For GSV the results are not significant. We use a daily panel data setting of 29 DJIA stocks over a period of three years. Our results are robust to various tests. They supports the idea that GSV and TV capture the beliefs of individual investors.

Our findings that TV has a higher impact on turnover and volatility than GSV is in line with the findings of Mao et al. (2015). While Twitter is a microblogging platform for news and opinions, GSV only measures what people search for if it raised their attention. Therefore it makes sense to distinguish between investor attention for GSV and investor sentiment for TV.

In chapter 4, I use GSV and TV as a proxy for investor attention and investor sentiment, to assess their predictive power on the RV of the DJIA in a time-series analysis over 2.5 years. Using a VAR model approach I find that GSV and TV Granger cause RV, controlling for macroeconomic and financial factors. Their impact is significant at the 1 % level for TV and at the 5 % level for GSV. First, I perform a one day ahead in-sample forecast of 222 days with an OLS regression. I find that models including GSV and TV improve the goodness of fit with an R^2 of 63.67 % compared to a classic AR (1) process with an R^2 of only 48.45 %. The impact of GSV and TV on the RV predictions is especially pronounced for the top 1 % RV. Second, I perform a one day ahead out-of-sample forecast for 334 days. Overall, GSV and TV do not lead to better predictions than the AR (1) process in terms of goodness of fit. While the R^2 of the AR (1) process is at 48.97 % it is only at 28.36 % for models including GSV and TV. In extreme situations GSV and TV can incorporate new information. For the top 1 % of RV and the lowest 1 % of RV the predictions including GSV and TV outperform a classical AR (1) model.

The robustness check with a shorter out-of-sample period improves these results. It underlines that the results of GSV and TV are not generally persistent but depend on the selected criteria.

5.2 Outlook

The focus of future research should be on improving the comprehension of investor beliefs. Four points can be derived from the previous chapters of this thesis. First, the question on how to measure GSV, TV and other platform and search volume data is important. The availability of data and its granularity is constantly improving and thus allows for further insights. With respect to Twitter, the granularity might help to improve the detection of specific mood types. Second, the use of different models using neural networks or machine learning approaches might detect connections better. Third, combining the measures of investor sentiment and investor attention could erase blind spots when it comes to the approximation of investor beliefs (e.g. Audrino et al., 2020). Fourth, this thesis shows that there is an impact of implied volatility, GSV and TV on economic factors (e.g. Mao et al., 2015). However, the results are not persistent and need further validation.

Bibliography

Abhyankar, A. H. (1995). Return and volatility dynamics in the FT-SE 100 stock index and stock index futures markets. *Journal of Futures Markets*, *15*(4), 457–488. https://doi.org/10.1002/fut.3990150405

Ahmed, S. I. (2018). Shorts on top as January Cboe bitcoin futures settle.

Alexander, R. M., & Gentry, J. K. (2014). Using social media to report financial results. *Business Horizons*, *57*(2), 161–167. https://doi.org/10.1016/j.bushor.2013.10.009

Alfarno, S., & Lux, T. (2007). A noise trader model as a generator of apparent financial power laws and long term memory. *Macroeconomic Dynamics*, *11*(S1). https://doi.org/10.1017/S1365100506060299

Amihud, Y., & Hurvich, C. M. (2004). Predictive Regressions: A Reduced-Bias Estimation Method. *Journal of Financial & Quantitative Analysis*, *39*(4), 813–841. https://doi.org/10.1017/S0022109000003227

Andersen, T. G., & Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review*, *39*(4), 885. https://doi.org/10.2307/2527343

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and Forecasting Realized Volatility. *Econometrica*, *71*(2), 579–625. https://doi.org/10.1111/1468-0262.00418

Andrei, D., & Hasler, M. (2015). Investor attention and stock market volatility. *Review of Financial Studies*, *28*(1), 33–72. https://doi.org/10.1093/rfs/hhu059

Antoniou, A., & Holmes, P. (1995). Futures trading, information and spot price volatility:

evidence for the FTSE-100 stock index futures contract using GARCH. *Journal of Banking & Finance*, *19*(1), 117–129. https://doi.org/10.1016/0378-4266(94)00059-C

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance*. https://doi.org/10.1111/j.1540-6261.2004.00662.x

Areal, N. M. P. C., & Taylor, S. J. (2002). The realized volatility of FTSE-100 futures prices. *Journal of Futures Markets*, *22*(7), 627–648. https://doi.org/10.1002/fut.10018

Arellano, M., & Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*. https://doi.org/10.2307/2297968

Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.05.010

Bain, B., & Mott, G. (2018, August 13). Can Elon Musk Tweet That? The SEC Is Digging In. Retrieved from https://www.bloomberg.com/news/articles/2018-08-07/can-elon-musk-tweet-that-the-sec-may-have-an-opinion-quicktake

Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, *61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. In *Journal of Economic Perspectives*. https://doi.org/10.1257/jep.21.2.129

Ballinari, D., & Behrendt, S. (2020). Structural breaks in online investor sentiment: A note on the nonstationarity of financial chatter. *Finance Research Letters*.

https://doi.org/10.1016/j.frl.2020.101479

Barber, Brad M; Odean, T. (2008). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *The Review of Financial Studies*, *21*(2), 785–818. https://doi.org/https://doi.org/10.1093/rfs/hhm079

Barber, B. M., Odean, T., & Zhu, N. (2009). Do retail trades move markets? *Review of Financial Studies*. https://doi.org/10.1093/rfs/hhn035

Barberis, N., & Huang, M. (2008). Stocks as Lotteries: The Implications of Probability Weighting for Security Prices. *American Economic Review*, *98*(5), 2066–2100. https://doi.org/10.1257/aer.98.5.2066

Barberis, N., Huang, M., & Thaler, R. H. (2006). Individual Preferences, Monetary Gambles, and Stock Market Participation: A Case for Narrow Framing. *American Economic Review*, *96*(4), 1069–1090. https://doi.org/10.1257/aer.96.4.1069

Barberis, N., & Thaler, R. (2003). *Financial Markets and Asset Pricing*. *Handbook of the Economics of Finance* (Vol. 1). https://doi.org/10.1016/S1574-0102(03)01027-6

Behrendt, S., & Schmidt, A. (2018). The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility. *Journal of Banking & Finance*, *96*, 355–367. https://doi.org/10.1016/j.jbankfin.2018.09.016

Belsley, D.A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics*. *Wiley Online Library*. New York.

Belsley, David A. (1991). A Guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, *4*(1), 33–50. https://doi.org/10.1007/BF00426854

Black, F. (1986). Noise. *The Journal of Finance*, *41*(3), 528–543.

https://doi.org/10.1111/j.1540-6261.1986.tb04513.x

Blume, M. E., & Friend, I. (1975). The Asset Structure of Individual Portfolios and Some Implications for Utility Functions. *The Journal of Finance*, *30*(2), 585. https://doi.org/10.2307/2978737

Boerse Stuttgart. (2018a). *Jaresbericht 2017*. Stuttgart. Retrieved from https://www.boerse-stuttgart.de/de/unternehmen/aktuelles/publikationen/jahresberichte/

Boerse Stuttgart. (2018b). Profile of Boerse Stuttgart. Retrieved June 4, 2018, from https://www.boerse-stuttgart.de/en/company/about-us/

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

Bologna, P., & Cavallo, L. (2002). Does the introduction of stock index futures effectively reduce stock market volatility? Is the "futures effect" immediate? Evidence from the Italian stock exchange using GARCH. *Applied Financial Economics*, *12*(3), 183–192. https://doi.org/10.1080/09603100110088085

Broihanne, M.-H., Merli, M., & Roger, P. (2016). Diversification, gambling and market forces. *Review of Quantitative Finance and Accounting*, *47*(1), 129–157. https://doi.org/10.1007/s11156-015-0497-1

Brooks, C. (1998). Predicting stock index volatility: can market volume help? *Journal of Forecasting*, *17*(1), 59–80. https://doi.org/10.1002/(SICI)1099-131X(199801)17

Brooks, C. (2014). *Introductory Econometrics for Finance 3rd edition*. *Cambridge University Press* (3rd ed.). https://doi.org/10.1017/CBO9781107415324.004

Brooks, C., Rew, A. G., & Ritson, S. (2001). A trading strategy based on the lead–lag

relationship between the spot index and futures contract for the FTSE 100. *International Journal of Forecasting*, *17*(1), 31–44. https://doi.org/10.1016/S0169-2070(00)00062-5

Brückner, R., Lehmann, P., Schmidt, M. H., & Stehle, R. (2015). Another German Fama/French Factor Data Set. *SSRN Electronic Journal*, 1–10. https://doi.org/10.2139/ssrn.2682617

Brunnermeier, M. K., Gollier, C., & Parker, J. A. (2007). Optimal Beliefs, Asset Prices, and the Preference for Skewed Returns. *American Economic Review*, *97*(2), 159–165. https://doi.org/10.1257/aer.97.2.159

Brunnermeier, M. K., & Parker, J. A. (2005). Optimal Expectations. *American Economic Review*, *95*(4), 1092–1118. https://doi.org/10.1257/0002828054825493

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, *52*(1), 57. https://doi.org/10.2307/2329556

Cboe. (2017). XBT-Cboe Bitcoin Futures. Retrieved June 4, 2018, from http://cfe.cboe.com/cfe-products/xbt-cboe-bitcoin-futures

Cboe. (2018). bitcoin-USD-Cboe-Futures Exchange. Retrieved June 4, 2018, from https://www.cftc.gov/dea/futures/deacboelf.htm

CFTC. (2017). CFTC Statement on Self-Certification of Bitcoin Products by CME, CFE and Cantor Exchange. Retrieved June 4, 2018, from https://www.cftc.gov/PressRoom/PressReleases/pr7654-17

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*(SUPPL.1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Commission, A. C. and C. (2017). *Targeting scams - Report of the ACCC on scams activity 2016*. Retrieved from https://www.accc.gov.au/publications/targeting-scams-report-on-

scam-activity/targeting-scams-report-of-the-accc-on-scam-activity-2016

Corwin, S. A., & Schultz, P. (2012). A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices. *Journal of Finance*, *67*(2), 719–760. https://doi.org/10.1111/j.1540-6261.2012.01729.x

D'Hondt, Catherine ; Roger, P. (2017). Investor sentiment and stock return predictability: The power of ignorance. *Finance*, *38*(2), 7–37.

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *The Journal of Finance*, *66*(5), 1461–1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

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. https://doi.org/10.1086/261703

Deutsche Börse AG. (2007). *Guide to Volatility Indices of Deutsche Börse*.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*. https://doi.org/10.2307/2286348

Dimpfl, T., & Jank, S. (2011). Can Internet search Queries help to predict stock market volatility ? *German Research*, (11), 1–32.

Dimpfl, T., & Jank, S. (2016). Can Internet Search Queries Help to Predict Stock Market Volatility? *European Financial Management*, *22*(2), 171–192. https://doi.org/10.1111/eufm.12058

Dimpfl, T., & Kleiman, V. (2017). Investor Pessimism and the German Stock Market: Exploring Google Search Queries. *German Economic Review*.

https://doi.org/10.1111/geer.12137

Dorn, D., Huberman, G., & Sengmueller, P. (2008). Correlated Trading and Returns. *The Journal of Finance*, *63*(2), 885–920. https://doi.org/10.1111/j.1540-6261.2008.01334.x

Driscoll, J. C., & Kraay, A. C. (1998). Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. *Review of Economics and Statistics*, *80*(4), 549–560. https://doi.org/10.1162/003465398557825

Easley, D., Kiefer, N. M., O'Hara, M., & Paperman, J. B. (1996). Liquidity, Information, and Infrequently Traded Stocks. *The Journal of Finance*, *51*(4), 1405–1436. https://doi.org/10.1111/j.1540-6261.1996.tb04074.x

Fama, E. F. (1970). Efficient Capital Markets, A Review of Theory and Empirical Work. *The Journal of Finance*. https://doi.org/10.2307/2329297

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Fink, C., & Johann, T. (2014). May I Have Your Attention , Please : The Market Microstructure of Investor Attention. *SSRN Working Paper*, (September), 59. https://doi.org/10.2139/ssrn.2139313

Gerth, M. (2017). Twittern und zittern. *WirtschaftsWoche*, 77–78.

Gillert, M. (2017). Marktmanipulation: Short-Attacken - Wie Anleger und Emittenten ins visier von Manipulatoren geraten. *BaFin Journal*, (5), 26–28. Retrieved from https://www.accc.gov.au/publications/targeting-scams-report-on-scam-activity/targeting-scams-report-of-the-accc-on-scam-activity-2016

Goetzmann, W. N., & Kumar, A. (2008). Equity Portfolio Diversification. *Review of Finance*, *12*(3), 433–463. https://doi.org/10.1093/rof/rfn005

Hamid, A., & Heiden, M. (2015). Forecasting volatility with empirical similarity and Google Trends. *Journal of Economic Behavior & Organization*, *117*, 62–81. https://doi.org/10.1016/j.jebo.2015.06.005

Hirshleifer, D. (2001). Investor psychology and asset pricing. *Journal of Finance*. https://doi.org/10.1111/0022-1082.00379

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*. https://doi.org/10.1016/j.jacceco.2003.10.002

Hoechle, D. (2007). Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal*. https://doi.org/The Stata Journal

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on Mathematics and Statistics*. https://doi.org/citeulike-article-id:2607115

Kahneman, D. (1973). *Attention and Effort* (Prentice-H). Prentice Hall.

Kahneman, D., & Tversky, A. (1979). Prospect Theory - An Analysis Of Decision Under Risk. *Econometrica*, *47*, 263–292. https://doi.org/10.2307/1914185

Kumar, A. (2007). Do the diversification choices of individual investors influence stock returns? *Journal of Financial Markets*, *10*(4), 362–390. https://doi.org/10.1016/j.finmar.2007.06.003

Kumar, A., & Lee, C. M. C. (2006). Retail Investor Sentiment and Return Comovements. *The*

Journal of Finance, *61*(5), 2451–2486. https://doi.org/10.1111/j.1540-6261.2006.01063.x

Lease, R. C., Lewellen, W. G., & Schlarbaum, G. G. (1974). The Individual Investor: Attributes and Attitudes. *The Journal of Finance*, *29*(2), 413–433. https://doi.org/10.1111/j.1540-6261.1974.tb03055.x

Leinweber, D. J., & Madhavan, A. N. (2001). Three Hundred Years of Stock Market Manipulations. *The Journal of Investing*, *10*(2), 1–10. https://doi.org/10.3905/joi.2001.319457

Lo, A. W., & Wang, J. (2000). Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory. *Review of Financial Studies*, *13*(2), 257–300. https://doi.org/10.1093/rfs/13.2.257

Long, J. B. De, Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise Trader Risk in Financial Markets Andrei Shleifer. *Journal of Political Economy*, *98*(4), 703–738. https://doi.org/10.1016/S0003-3472(05)80894-2

Lux, T., & Marchesi, M. (1999). Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, *397*(6719), 498–500. https://doi.org/10.1038/17290

Mao, H., Counts, S., & Bollen, J. (2015). Quantifying the effects of online bullishness on international financial markets. *ECB Workshop on Using Big Data for Forecasting and Statistics*, 1–15.

Mehlhorn, M. (2018). *Marktmikrostruktur und die aktienspezifische Aufmerksamkeit der Marktteilnehmer aus theoretischer und empirischer Sicht* (1st ed.). Dr. Kovac. Retrieved from https://www.amazon.de/Marktmikrostruktur-aktienspezifische-Aufmerksamkeit-Marktteilnehmer-Schriftenreihe/dp/383009728X/ref=sr_1_1?ie=UTF8&qid=1519482324&sr=8-

1&keywords=mehlhorn+marc

Mitton, T., & Vorkink, K. (2007). Equilibrium Underdiversification and the Preference for Skewness. *Review of Financial Studies*, *20*(4), 1255–1288. https://doi.org/10.1093/revfin/hhm011

Odean, T. (1998). Are Investors Reluctant to Realize Their Losses? *The Journal of Finance*, *53*(5), 1775–1798. https://doi.org/10.1111/0022-1082.00072

Odean, T. (1999). Do Investors Trade Too Much? *American Economic Review*, *89*(5), 1279–1298. https://doi.org/10.1257/aer.89.5.1279

Oliveira, N., Cortez, P., & Areal, N. (2013). On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 355–365). https://doi.org/10.1007/978-3-642-40669-0_31

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125–144. https://doi.org/10.1016/j.eswa.2016.12.036

Osipovich, A. (2018, January 7). Little Guys and Big Trading Firms Square Off in Bitcoin Futures Arena. *The Wall Street Journal*.

Peter, G., Darko, A., Igor, M., & Miha, G. (2017). Twitter sentiment around the earnings announcement events. *PLoS ONE*, *12*(2). https://doi.org/10.1371/journal.pone.0173151

Poteshman, A. M. (2001). Underreaction, Overreaction, and Increasing Misreaction to Information in the Options Market. *The Journal of Finance*, *56*(3), 851–876. https://doi.org/10.1111/0022-1082.00348

Roger, P. (2014). The 99% Market Sentiment Index. *Finance*, (35), 53–96.

Roodman, D. (2006). How to do Xtabond2: An Introduction to Difference and System GMM in Stata. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.982943

Schneller, D., Heiden, S., Heiden, M., & Hamid, A. (2018). Home is Where You Know Your Volatility – Local Investor Sentiment and Stock Market Volatility. *German Economic Review*. https://doi.org/10.1111/geer.12125

SEC. (2013). *Report of Investigation Pursuant to Section 21(a) of the Securities Exchange Act of 1934: Netflix, Inc., and Reed Hastings*. Retrieved from https://www.sec.gov/litigation/investreport/34-69279.htm

SEC. (2014). Updated Investor Alert: Social Media and Investing - Avoiding Fraud. Retrieved November 29, 2017, from https://www.investor.gov/additional-resources/news-alerts/alerts-bulletins/updated-investor-alert-social-media-investing

SEC. (2015). Updated Investor Alert: Social Media and Investing -- Stock Rumors. Retrieved November 29, 2017, from https://www.sec.gov/oiea/investor-alerts-bulletins/ia_rumors.html

See-To, E. W. K., & Yang, Y. (2017). Market sentiment dispersion and its effects on stock return and volatility. *Electronic Markets*, *27*(3), 283–296. https://doi.org/10.1007/s12525-017-0254-5

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

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

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*. https://doi.org/10.2307/1912017

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*. https://doi.org/10.1111/j.1468-036X.2013.12007.x

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3), 375–421. https://doi.org/10.1016/S0304-405X(99)00041-0

Tafti, A., Zotti, R., & Jank, W. (2016). Real-time diffusion of information on twitter and the financial markets. *PLoS ONE*, *11*(8). https://doi.org/10.1371/journal.pone.0159226

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. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Verma, R., & Verma, P. (2007). Noise trading and stock market volatility. *Journal of Multinational Financial Management*, *17*(3), 231–243. https://doi.org/10.1016/j.mulfin.2006.10.003

Vinet, L., & Zhedanov, A. (2011). A 'missing' family of classical orthogonal polynomials. *Journal of Physics A: Mathematical and Theoretical*, *44*(8), 085201. https://doi.org/10.1088/1751-8113/44/8/085201

Vlastakis, N., & Markellos, R. N. (2012). Information demand and stock market volatility. *Journal of Banking and Finance*, *36*(6), 1808–1821. https://doi.org/10.1016/j.jbankfin.2012.02.007

Vozlyublennaia, N. (2014). Investor attention, index performance, and return predictability. *Journal of Banking and Finance*, *41*(1), 17–35. https://doi.org/10.1016/j.jbankfin.2013.12.010

Wang, Y.-H. H., Keswani, A., & Taylor, S. J. (2006). The relationships between sentiment, returns and volatility. *International Journal of Forecasting*, *22*(1), 109–123. https://doi.org/10.1016/j.ijforecast.2005.04.019

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*. https://doi.org/10.2307/1912934

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear." *Procedia - Social and Behavioral Sciences*. https://doi.org/10.1016/j.sbspro.2011.10.562

Appendix

Appendix 1: Stocks in the sample

Ticker	Availability of data Google	Twitter	Volatility	Number of Tweets (total)	Google Volume (total)	Number of Tweets (av. per day)	Google Volume (av. per day)
AAPL				2,647,627	21,148	2,648	12
ARNC	NA	NA	NA				
AXP				84,935	136,708	85	75
BA				142,294	16,549	142	9
BAC	NA	NA	NA				
CAT				130,045	52,844	130	29
CSCO				185,010	48,743	185	27
DVX				129,906	63,359	130	35
DD				59,013	52,531	59	29
DIS				245,473	34,618	245	19
GE				205,987	90,682	206	50
GS				362,825	18,332	362	10
HD				108,580	78,680	108	43
HPQ	NA	NA	NA				
IBM				278,778	63,948	278	35
INTC				226,605	51,244	226	28
JMP			NA				
JNJ				151,564	50,425	151	28
KO				140,081	38,022	140	21
MCD				172,928	77,517	173	42
MMM				55,000	61,790	55	34
MRK				137,731	51,214	138	28
MSFT				487,582	48,611	487	27
NKE				139,963	77,318	140	42
PFE				157,539	33,763	158	18
PG				85,388	18,393	85	10
T				242,827	101,652	242	56
TRV				26,919	78,757	27	43
UNHP			NA				
UTX				46,606	25,205	47	14
V				233,727	129,584	233	71
VZ				138,472	69,972	138	38
WMT				200,951	86,319	201	47
XOM				201,932	55,330	202	30
Total	31	31	29	7,426,288	1,733,258	7,419	949

Overview on the stocks of the DJIA in the sample. TV in total and on average per day. GSV in total and on average per day. For the time period from 6 June 2013 until 31 December 2016. Data is available for 29 stocks of the DJIA.

Appendix 2: Summary statistic – spread per stock

Company	Mean	Std. Dev.	Min	Max	N
AAPL	0.0103782	0.0116204	0	0.083484	872
AXP	0.0092676	0.0107487	0	0.1434309	872
BA	0.0096986	0.0117847	0	0.0992901	872
CAT	0.0109405	0.0115995	0	0.0711514	872
CSCO	0.0096921	0.0090418	0	0.0581376	872
CVX	0.0104332	0.0110567	0	0.0774814	872
DD	0.0101192	0.0114946	0	0.103698	872
DIS	0.0087486	0.0098998	0	0.0751401	872
GE	0.0088762	0.009651	0	0.0929266	872
GS	0.0101573	0.0111224	0	0.0802625	872
HD	0.0093221	0.0100226	0	0.0761029	872
IBM	0.0087695	0.0091146	0	0.0798768	872
INTC	0.0103975	0.0106493	0	0.0742689	872
JNJ	0.0073716	0.0080397	0	0.0736634	872
KO	0.0074992	0.0073212	0	0.0572615	872
MCD	0.0075277	0.0086941	0	0.1105769	872
MMM	0.0070537	0.0077085	0	0.0560245	872
MRK	0.0099797	0.0104666	0	0.0783858	872
MSFT	0.0103603	0.0115216	0	0.0716006	872
NKE	0.0097861	0.0107593	0	0.0866509	872
PFE	0.0103988	0.0109323	0	0.0915839	872
PG	0.0074801	0.0077683	0	0.0590997	872
T	0.008157	0.0088587	0	0.069135	872
TRV	0.0070804	0.0078785	0	0.0596684	872
UTX	0.0087653	0.0096453	0	0.0741878	872
V	0.0099096	0.0127093	0	0.1883466	872
VZ	0.0085925	0.008744	0	0.062108	872
WMT	0.0086108	0.011663	0	0.2324585	872
XOM	0.0091261	0.0089744	0	0.0618544	872
TOTAL	0.0091207	0.0101459	0	0.2324585	25288

Overview on the 29 stocks of the DJIA in the sample. For each stock we report the descriptive statistic of the zero spread. Calculated following Corwin and Schultz (2012). All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA.

Appendix 3: Market entry – per sector

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Basic Materials[63]	Communication Servics	Consumer Cyclical	Consumer Defensive	Energy	Financial Services	Healthcare	Industrials	Technology
VARIABLES	$\Delta Turnover_t$	$\Delta Turnover_t$	$\Delta Turnover_t$	$\Delta Turnover_t$	$\Delta Turnover_t$	$\Delta Turnover_t$	$\Delta Turnover_t$	$\Delta Turnover_t$	$\Delta Turnover_t$
$\Delta Turnover_{t-1}$	-0.210***	-0.373***	-0.322***	-0.313***	-0.376***	-0.345***	-0.380***	-0.317***	-0.366***
	(-1.38e+13)	(-8.238)	(-15.558)	(-12.766)	(-9.783)	(-17.249)	(-13.634)	(-16.754)	(-19.965)
$\Delta Squared\_Return_t$	-0.019***	0.028***	0.028***	0.031***	0.014***	0.029***	0.026***	0.031***	0.032***
	(-3.18e+12)	(5.123)	(10.529)	(9.615)	(4.278)	(9.788)	(7.920)	(11.517)	(14.365)
$\Delta Squared\_Return_{t-1}$	0.034***	0.011*	0.012***	0.012***	0.004	0.016***	0.017***	0.019***	0.017***
	(6.18e+12)	(2.266)	(5.326)	(3.897)	(0.972)	(6.109)	(6.024)	(7.742)	(9.132)
ΔGSV_t	-0.245***	0.030	0.127**	-0.012	-0.004	0.038	0.029	0.063**	0.137***
	(-1.07e+13)	(0.587)	(3.272)	(-0.608)	(-0.127)	(1.132)	(1.823)	(3.194)	(6.111)
ΔGSV_{t-1}	-0.240***	0.020	0.086*	-0.001	0.006	0.012	0.019	0.030	0.054*
	(-8.87e+12)	(0.345)	(2.242)	(-0.058)	(0.200)	(0.335)	(1.222)	(1.700)	(2.411)
ΔTV_t	0.643***	0.121***	0.196***	0.140***	0.071***	0.132***	0.129***	0.197***	0.228***
	(1.98e+13)	(4.051)	(13.240)	(7.384)	(3.308)	(7.282)	(8.064)	(10.618)	(14.381)
ΔTV_{t-1}	0.052***	0.089***	0.101***	0.066***	0.040*	0.064***	0.055***	0.088***	0.163***
	(3.62e+12)	(4.203)	(8.811)	(4.392)	(1.987)	(4.393)	(4.318)	(7.979)	(10.743)
Constant	0.241***	-0.093	-0.260***	-0.167***	-0.233***	0.255***	-0.071	-0.010	-0.932***
	(1.14e+13)	(-1.669)	(-27.439)	(-6.335)	(-7.871)	(13.280)	(-1.757)	(-0.501)	(-17.693)
Time Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES	YES
Company Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES	YES
R-squared	1.000	0.866	0.655	0.728	0.876	0.659	0.739	0.671	0.698
N	865	1,760	3,553	2,560	1,772	3,553	2,604	3,757	4,377

Column (1) to (9) report the result on market entry: $\text{Turnover}_{it} = \alpha + \beta_1 \text{Turnover}_{it-1} + \beta_2 \text{SquaredReturn}_{it} + \beta_3 \text{SquaredReturn}_{it-1} + \beta_4 \text{GSV}_{it} + \beta_5 \text{GSV}_{it-1} + \beta_6 \text{TV}_{it} + \beta_7 \text{TV}_{i,t-1} + \mu_i + \lambda_t + \nu_{it}$. Column (1) reports the results for the sector Basic Materials. Column (2) stands for the sector Communication Services. Column (3) stands for the sector Consumer Cyclical. Column (4) stands for the sector Consumer Defensive. Column (5) stands for the sector Energy. Column (6) stands for the sector Financial Services. Column (7) stands for the sector Healthcare. Column (8) stands for the sector Industrials and column (9) for the sector Technology. The dependent variable ΔTurnover stands for the percentage changes in trading activity. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV to proxy investor attention and investor sentiment. ΔSquared_Return stands for the percentage change in squared returns as a proxy news.

[63] Only one company

All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. All regressions are panel fixed effects regression, including time and company fixed effects. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

Appendix 4: Share of noise trader – per sector

VARIABLES	(1) Basic Materials Volatility$_t$	(2) Communication Servics Volatility$_t$	(3) Consumer Cyclical Volatility$_t$	(4) Consumer Defensive Volatility$_t$	(5) Energy Volatility$_t$	(6) Financial Services Volatility$_t$	(7) Healthcare Volatility$_t$	(8) Industrials Volatility$_t$	(9) Technology Volatility$_t$
ΔVolatility$_{t-1}$	-0.169***	-0.482***	-0.451***	-0.489***	-0.445***	-0.482***	-0.478***	-0.501***	-0.443***
	(-1.98e+12)	(-12.745)	(-26.322)	(-20.017)	(-12.512)	(-25.126)	(-23.527)	(-30.545)	(-29.440)
ΔTurnover$_t$	0.094***	0.531***	0.508***	0.515***	0.511***	0.484***	0.528***	0.572***	0.603***
	(7.93e+11)	(5.903)	(12.531)	(9.482)	(5.522)	(14.953)	(9.321)	(18.290)	(18.730)
ΔTurnover$_{t-1}$	0.091***	0.275***	0.207***	0.234***	0.126	0.222***	0.233***	0.230***	0.222***
	(6.65e+11)	(3.311)	(5.567)	(4.990)	(1.239)	(7.573)	(4.658)	(7.738)	(6.673)
ΔSquared_Return$_t$	-0.029***	0.047***	0.048***	0.045***	0.044***	0.050***	0.055***	0.036***	0.042***
	(-2.41e+12)	(5.892)	(14.627)	(9.324)	(6.420)	(13.658)	(9.943)	(10.435)	(15.913)
ΔSquared_Return$_{t-1}$	-0.014***	0.016*	0.020***	0.023***	0.022*	0.022***	0.022***	0.019***	0.023***
	(-1.88e+12)	(2.195)	(6.731)	(4.648)	(2.554)	(6.235)	(4.519)	(5.931)	(7.703)
ΔGSV$_t$	0.004***	-0.098	-0.103*	0.007	-0.014	0.019	-0.024	0.020	-0.038
	(9.67e+10)	(-1.447)	(-2.289)	(0.271)	(-0.272)	(0.409)	(-1.086)	(1.002)	(-1.379)
ΔGSV$_{t-1}$	0.218***	-0.013	-0.040	-0.006	-0.002	0.051	0.003	0.003	-0.010
	(3.88e+12)	(-0.162)	(-0.904)	(-0.196)	(-0.038)	(1.179)	(0.126)	(0.178)	(-0.320)
ΔTV$_t$	0.424***	0.019	0.049**	0.030	0.063	0.052***	0.016	0.070***	0.007
	(5.02e+12)	(0.611)	(2.868)	(1.478)	(1.800)	(3.330)	(0.768)	(4.373)	(0.474)
ΔTV$_{t-1}$	-0.089***	0.013	0.034	0.021	0.027	0.032*	0.009	0.051***	-0.015
	(-1.62e+12)	(0.364)	(1.885)	(1.013)	(0.888)	(2.163)	(0.468)	(3.583)	(-0.825)
Constant	-0.079***	0.329***	0.218***	-0.164***	-0.135*	0.106***	0.047	-0.032	-0.273***
	(-2.47e+12)	(3.938)	(12.713)	(-3.348)	(-2.271)	(3.892)	(0.851)	(-1.582)	(-4.584)
Time Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES	YES
Company Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES	YES
R-squared	1.000	0.879	0.689	0.739	0.856	0.721	0.762	0.726	0.667
N	865	1760	3553	2560	1772	3553	2604	3757	4377

Column (1) to (9) report the result on market entry: $\text{Volatility}_{it} = \alpha + \beta_1\text{Volatility}_{it-1} + \beta_2\text{SquaredReturn}_{it} + \beta_3\text{SquaredReturn}_{it-1} + \beta_4\text{Turnover}_{it} + \beta_5\text{Turnover}_{it-1} + \beta_6\text{GSV}_{it} + \beta_7\text{GSV}_{it-1} + \beta_8\text{TV}_{it} + \beta_9\text{TV}_{it-1} + \mu_i + \lambda_t + \nu_{it}$. Column (1) reports the results for the sector Basic Materials. Column20 (2) stands for the sector Communication Services. Column (3) stands for the sector Consumer Cyclical. Column (4) stands for the sector Consumer Defensive. Column (5) stands for the sector Energy. Column (6) stands for the sector Financial Services. Column (7) stands for the sector Healthcare. Column (8) stands for the sector Industrials and column (9) for the sector Technology. The dependent variable ΔVolatility stands for the percentage changes in volatility. The control variables ΔGSV and ΔTV stand for percentage changes in GSV and TV. ΔSquared_Return stands for the percentage change in squared returns as a proxy news. ΔTurnover stands for the percentage changes in trading activity. All variables are calculated on a daily basis from 6 June 2013 until 31 December 2016 for 29 stocks of the DJIA. All regressions are panel fixed effects regression, including time and company fixed effects. The regressions are estimated with Driscoll and Kraay standard errors. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

Appendix 5: Descriptive statistic - original variables

Panel A: Correlation Analysis							
	RV	GSV	TV	VXD	Return	Futures	T3
RV	1.0000						
GSV	0.3829***	1.0000					
TV	0.3789***	0.5470***	1.0000				
VXD	0.7519***	0.4617***	0.4069***	1.0000			
Return	-0.2028***	-0.0300	-0.1945***	-0.1926***	1.0000		
Futures	0.5981***	0.3514***	0.3403***	0.5827***	-0.2515***	1.0000	
T3	-0.0843**	0.2217***	-0.1601***	0.0295	0.0415	0.0653*	1.0000
Panel B: Summary Statistics							
	Mean	S.D.	Skewness	Kurtosis	Min.	Max.	N
RV	-5.37	0.46	0.30	3.18	-6.76	-3.48	668
GSV	5.48	4.76	13.02	240.81	2.00	100.00	668
TV	65.08	53.60	5.51	57.32	8.05	779.00	668
VXD	15.14	3.60	1.67	6.56	10.18	34.51	668
Return	0.00	0.01	-0.27	4.98	-0.04	0.04	667
Futures	161000	73019	1.87	9.18	15402	653000	668
T3	0.15	0.15	0.77	2.36	-0.02	0.54	668

Panel A: Correlation reports the correlation of the variables log_RV, GSV, TV, implied volatility of the DJIA (VXD), 3-Month Treasury Bill (T3), futures on the DJIA and the daily return of the DJIA. Panel B reports the summary statistics. All variables are calculated on a daily basis from 5 January 2014 until 30 December 2016 for DJIA. *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

Appendix 6: VAR 2 lags

	Model 1			Model 2		
	(1)	(2)	(3)	(1)	(2)	(3)
	RV_t	Gr_t	$T2r_t$	RV_t	Tr_t	$G2r_t$
RV_{t-1}	0.57105***	0.31029	3.19795	0.57105***	5.05597	0.08617
	(14.86755)	(0.63954)	(0.69333)	(14.86755)	(0.94055)	(0.20701)
RV_{t-2}	0.18991***	-0.67610	-6.00023	0.18991***	-10.04877	-0.23066
	(4.96663)	(-1.39977)	(-1.30671)	(4.96663)	(-1.87771)	(-0.55662)
Gr_{t-1}	-0.00642*	0.13601***	-0.03398			
	(-2.08456)	(3.49711)	(-0.09191)			
Gr_{t-2}	-0.00454	0.05785	0.09683			
	(-1.47520)	(1.48699)	(0.26178)			
$T2r_{t-1}$	-0.00067*	0.00395	0.26986***			
	(-2.08180)	(0.97564)	(7.01955)			
$T2r_{t-2}$	-0.00014	0.00503	0.14127***			
	(-0.44042)	(1.23396)	(3.64801)			
Tr_{t-1}				-0.00077**	0.25017***	-0.00216
				(-2.81250)	(6.49500)	(-0.72513)
Tr_{t-2}				-0.00031	0.14553***	-0.00019
				(-1.10566)	(3.76032)	(-0.06482)
$G2r_{t-1}$				-0.00243	-0.97694	0.15569***
				(-0.67794)	(-1.94959)	(4.01234)
$G2r_{t-2}$				-0.00369	-0.58288	0.05359
				(-1.02827)	(-1.15955)	(1.37687)
$R3_t$	-0.08559	-0.14900	-0.50699	-0.08559	-1.39918	-0.08697
	(-1.02004)	(-0.14058)	(-0.05032)	(-1.02004)	(-0.11915)	(-0.09564)
Constant	-1.28369***	-1.96322	-14.98638	-1.28369***	-26.74226	-0.77779
	(-8.21601)	(-0.99474)	(-0.79873)	(-8.21601)	(-1.22295)	(-0.45935)
Observations	665	665	665	665	665	665
RMSE	0.32	4.03	38.30	0.32	44.64	3.46
LL	-180	-1870	-3360	-180	-3470	-1760
R^2	0.53	0.03	0.12	0.53	0.12	0.03
AIC	16.32	16.32	16.32	16.32	16.32	16.32

For model 1 and 2 the table reports the results of a basic VAR regression with 2 lags. For model 1: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 RV_{t-2} + \beta_3 Gr_{t-1} + \beta_4 Gr_{t-2} + \beta_5 T2r_{t-1} + \beta_6 T2r_{t-2} + \beta_7 R3_t + u_t$ for model 2: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 RV_{t-2} + \beta_3 Tr_{t-1} + \beta_4 Tr_{t-2} + \beta_5 G2r_{t-1} + \beta_6 G2r_{t-2} + \beta_7 R3_t + u_t$. Column (1) reports results for log RV (RV). Column (2) reports for model 1 (model 2) the result for the residuals of GSV (TV) Gr_t (Tr_t). Column (3) reports for model 1 (model 2) the result for the residuals of TV (GSV) $T2r_t$ ($G2r_t$). As exogenous variable all regressions include R3, a combined residual of the macroeconomic and financial factor variables. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL). *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.

Appendix 7: VAR 3 lags

	Model 1			Model 2		
	(1)	(2)	(3)	(1)	(2)	(3)
	log_RV	Gr	T2r	log_RV	Tr	G2r
RV_{t-1}	0.54680***	0.31586	4.22390	0.54680***	6.11527	0.04478
	(14.04165)	(0.63690)	(0.90824)	(14.04165)	(1.12220)	(0.10575)
RV_{t-2}	0.12264**	-0.65757	-2.85480	0.12264**	-6.79236	-0.35648
	(2.78662)	(-1.17317)	(-0.54313)	(2.78662)	(-1.10285)	(-0.74488)
RV_{t-3}	0.12094**	-0.05092	-4.14581	0.12094**	-4.45071	0.14637
	(3.11283)	(-0.10291)	(-0.89348)	(3.11283)	(-0.81860)	(0.34647)
Gr_{t-1}	-0.00581	0.13682***	-0.13148			
	(-1.89056)	(3.49595)	(-0.35825)			
Gr_{t-2}	-0.00408	0.05880	-0.00406			
	(-1.31618)	(1.49058)	(-0.01098)			
Gr_{t-3}	-0.00122	0.00158	0.00777			
	(-0.39892)	(0.04054)	(0.02121)			
$T2r_{t-1}$	-0.00065*	0.00426	0.24479***			
	(-2.02621)	(1.04097)	(6.37463)			
$T2r_{t-2}$	-0.00021	0.00572	0.09778*			
	(-0.61867)	(1.35503)	(2.47082)			
$T2r_{t-3}$	0.00022	-0.00258	0.16521***			
	(0.67708)	(-0.62318)	(4.26182)			
Tr_{t-1}				-0.00074**	0.22905***	-0.00096
				(-2.65404)	(5.90153)	(-0.31746)
Tr_{t-2}				-0.00033	0.11241**	0.00182
				(-1.16499)	(2.82446)	(0.58992)
Tr_{t-3}				0.00011	0.11079**	-0.00673*
				(0.38513)	(2.84516)	(-2.22530)
$G2r_{t-1}$				-0.00191	-0.93088	0.15256***
				(-0.53602)	(-1.86840)	(3.94064)
$G2r_{t-2}$				-0.00285	-0.44252	0.04418
				(-0.78963)	(-0.87630)	(1.12584)
$G2r_{t-3}$				-0.00254	-0.87965	0.05600
				(-0.70960)	(-1.75588)	(1.43869)
$R3_t$	-0.08234	-0.17218	0.23466	-0.08234	-0.79634	-0.13688
	(-0.98783)	(-0.16220)	(0.02357)	(-0.98783)	(-0.06827)	(-0.15102)
Constant	-1.12546***	-2.10678	-14.77540	-1.12546***	-27.39090	-0.89259
	(-6.92103)	(-1.01729)	(-0.76081)	(-6.92103)	(-1.20368)	(-0.50480)
Observations	664	664	664	664	664	664
RMSE	0.32	4.04	37.88	0.32	44.39	3.45
LL	-174	-1860	-3350	-174	-3460	-1760
R^2	0.54	0.03	0.14	0.54	0.14	0.04
AIC	16.31	16.31	16.31	16.31	16.31	16.31

For model 1 and 2 the table reports the results of a basic VAR regression with 3 lags. For model 1: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 RV_{t-2} + \beta_3 RV_{t-3} + \beta_4 Gr_{t-1} + \beta_5 Gr_{t-2} + \beta_6 Gr_{t-3} + \beta_7 T2r_{t-1} + \beta_8 T2r_{t-2} + \beta_9 T2r_{t-3} + \beta_{10} R3_t + u_t$ for model 2: $RV_t = \alpha + \beta_1 RV_{t-1} + \beta_2 RV_{t-2} + \beta_3 RV_{t-3} + \beta_4 Tr_{t-1} + \beta_5 Tr_{t-2} + \beta_6 Tr_{t-3} + \beta_7 G2r_{t-1} + \beta_8 G2r_{t-2} + \beta_9 G2r_{t-3} + \beta_{10} R3_t + u_t$ Column (1) reports results for log RV (RV). Column (2) reports for model 1 (model 2) the result for the residuals of GSV (TV) Gr_t (Tr_t). Column (3) reports for model 1 (model 2) the result for the residuals of TV (GSV) $T2r_t$ ($G2r_t$). As exogenous variable all regressions include R3, a combined residual of the macroeconomic and financial factor variables. The forecasting quality is measured by the root mean squared error (RMSE) and the goodness of fit by R^2 and log likelihood (LL). *, **, *** indicate statistical significance at the 10 %, 5 % or 1 % levels, respectively.