Alexander Mesel Isom

Emil Jenssen

Volatility forecasting with Twitter message volume

Improving forecasting ability of the Heterogenous Autoregressive model of Realized Volatility

Master's thesis spring 2020
Oslo Business School
Oslo Metropolitan University
Master's program in business administration

Abstract

We test the predictive power of Twitter message volume on realized volatility in a panel of 22 companies from S&P 100. The relationship is tested using OLS regression, with Newey-West standard errors. The forecasting ability of the two best performing Twitter variables is tested in models inspired by the Heterogenous Autoregressive model of Realized Volatility. We employ the Two-Scales Estimator (TSE) to achieve more precise estimates of realized volatility. TSE takes advantage of high-frequency return data, while simultaneously correcting for microstructure noise.

Findings indicate a relationship between Twitter message volume and realized volatility, made most explicit by employing variables that only contain messages accumulated outside of trading hours. Twitter messages containing cashtags were found to hold greater predictive power compared to messages containing company name mentions. However, a model containing both Twitter variables were superior in terms of forecasting ability. Company name messages were found to explain systematic risk, while cashtag messages were found to explain idiosyncratic risk.

Preface

The basis for this research was our interest in studying the value of social media content as a financial predictor. Twitter is one of the most popular social media channels and has an easily accessible API. Therefore, it made sense to consider it for our research purposes. Coupled with our interest in risk management, we quickly turned our gaze towards volatility modeling. The importance of volatility modeling is illustrated by how often we encounter crisis in the stock market. An important component for improving volatility models is finding exogenous variables that improve predictive ability. As such, the idea of improving volatility models with Twitter message data was born.

This thesis is the final work of our Master's degree in Business Administration at Oslo Metropolitan University. The work has been challenging and rewarding. We would like to take this opportunity to thank our supervisor, Associate Professor Einar Belsom. He provided valuable guidance, constructive criticism and showed genuine interest in our work, as well as lifting our spirits when needed.

Oslo, June 15th 2020.

Alexander M. Isom

Alexander Mesel Isom

Emil Jenssen

Emil Jenssen

Table of contents

1 Introduction

Since the early 2000s, social media and blogging have skyrocketed in popularity and become global phenomena. Because of this, we now have more information available to us than ever before in terms of ideas and opinions. Our digital property is in high demand, and many are eagerly data mining social media sites to uncover valuable information that may be used for commercial purposes. Sites like Twitter, with an estimated monthly user base of 330 million (Twitter Inc, 2019a, p. 5), is a popular microblogging website that is utilized for this. Using a few lines of code, one can gain access to Twitter's free application programming interface (API) that contains vast amounts of information about users and their messages. Due to its short message system, low cost of extraction, substantive user base and posting frequency, Twitter has also become a valuable data source for academic purposes.

Many people, including investors, use Twitter to stay updated on news and trends, and to share opinions on certain topics. In the context of behavioral finance, Twitter provides useful insight to communication that affects investor behavior and how it relates to markets. This recent field of study, where the effect of microblogging on stock markets is measured, use social media data from sites like Twitter to predict variables such as return, volatility and trading volume.

Daily volatility modeling has in recent years been greatly improved by the development of robust estimators that harness the power of high-frequency intraday data. Volatility forecasting has many applications in the field of finance, including security pricing and hedging, market making and risk management. Earlier works regarding volatility modeling and forecasting, with the inclusion of social media data, show promising results. A paper from Antweiler and Frank (2004) tried to predict market volatility using messages from stock message boards. Results suggested that stock messages had predictive power and that they could help forecast volatility both daily and within the trading day, using two different volatility models. Similarly, Dimpfl and Jank (2016) tried to predict the Dow Jones realized volatility by adding internet search queries of its name to an autoregressive volatility model. They found evidence that the inclusion of search queries improved the forecasting ability of the model, and that there existed a relationship between search queries from the previous day and realized volatility on the subsequent day.

In regard to Twitter, Sprenger et al. (2014), Tafti et al. (2016) and Oliveira et al. (2013) all found a link between posting volume and trading volume, which could lend support to the argument that a relationship between volatility and Twitter exists, because of how liquidity affects volatility. However, Oliveira et al. (2017) put this into question, as a more comprehensive and robust model failed to prove that Twitter message volume could improve volatility forecasts significantly. They did find a connection between posting volume and trading volume, but only in conjunction with sentiment indicators. Further, they conclude that the complimentary value of different internet data sources is unclear, and that other combinations might provide better financial predictions. Another recent study conducted, by Behrendt and Schmidt (2018), estimated intraday volatility for a panel consisting of companies from the Dow Jones Industrial Average. While some co-movements between Twitter message volume, sentiment and volatility were found, they conclude that high-frequency data from Twitter is not useful when forecasting intraday volatility.

Approaching volatility modeling with a daily perspective, we employ the Heterogenous Autoregressive model of Realized Volatility (HAR-RV) (Corsi, 2009) in a panel data setup. To achieve more precise estimates of realized volatility, we apply the Two-Scales Estimator (Zhang et al., 2005) with high-frequency 1-minute return data. This provides the foundational framework for our analysis. For a panel consisting of 22 randomly selected companies from S&P 100, we gather Twitter message volume for cashtag and company name mentions to create attention indicators. To our knowledge, our study is the first of its kind to employ the HAR-RV model with panel data, in conjunction with Twitter message volume, to model daily volatility. Our two main objectives with this paper are: to study the effect of attention on volatility, using Twitter message volume accumulated prior to trading hours, and use Twitter data to improve volatility forecasts. We formalize these objectives into three hypotheses:

Hypothesis 1: *Changes in company attention on Twitter, measured by message volume related to company stock accumulated prior to trading hours, is associated with changes in volatility.*

Hypothesis 2: *Changes in company attention on Twitter, measured by message volume related to the company in general accumulated prior to trading hours, is associated with changes in volatility.*

Hypothesis 3: *Company attention can be utilized to improve the forecasting ability of volatility models.*

The remaining sections of this paper are organized as follows. Section 2 outlines relevant volatility theory and data, including a presentation of the HAR-RV model from Corsi (2009), as well as detailed information regarding our chosen volatility estimator, and sample statistics. Section 3 describes relevant Twitter theory, our data collection process, variable selection, as well as qualitative and quantitative assessments of the sample. Section 4 connects volatility and Twitter in the panel we employ for our empirical study, and presents the baseline HAR-RV model for performance comparison. Section 5 describes our empirical approach, where we specify our models and tests. Section 6 contains results and discussion, where we test our hypotheses and discuss our findings. Section 7 concludes.

2 Volatility

In this section, we will introduce the HAR-RV model (Corsi, 2009). First, we must look at the framework that the model is derived from, and how to measure realized volatility.

2.1 Theory

In a review paper on realized volatility by McAleer and Medeiros (2008), and similarly in Corsi (2009) and Andersen et al. (2001), integrated variance on day t is defined as the integral of the instantaneous variance of the one-day interval $[t - 1d\,,\ t]$ of a continuous time diffusion process for the logarithmic prices of an asset:

$$IV_t = \int_{t-1d}^{t} \sigma^2(\omega)d\omega \tag{1}$$

Although processes behind asset prices are assumed to be continuous, the financial markets are inherently discrete. Asset prices change due to transactions occurring in time intervals that are not of equal length, therefore, the underlying integrated volatility of asset returns is unobservable. However, Andersen et al. (2001, p. 42) showed that integrated volatility can be estimated by the realized volatility measure: "By sampling intraday returns sufficiently frequently, the realized volatility can be made arbitrarily close to the underlying integrated volatility, the integral of instantaneous volatility over the interval of interest, which is a natural volatility measure."

Realized volatility on day t is defined as the square root of the sum of intraday squared log returns:

$$RV_t = \sqrt{\sum_{i=1}^{M-1} r_{t,i}^2} \tag{2}$$

where M is the number of equally spaced observations of intraday prices, and $r_{t,i}$ is the continuously compounded return for the time interval $\left[i, i + \frac{1}{M}\right]$ for day t (Corsi, 2009).

Furthermore, Andersen et al. (2003) showed that a simple vector autoregressive realized volatility model systematically outperformed other volatility models, like GARCH and FIEGARCH, in out-of-sample forecasting. Realized volatility provides a more precise estimate of the current volatility because it makes use of valuable intraday information. A superior estimate of today's volatility should yield a superior forecast of tomorrow's volatility, they argue.

2.1.1 HAR-RV

HAR-RV is motivated by the work of Müller, et al. (1993) on the Heterogenous Market Hypothesis. Müller, et al. looked at the properties of volatility and proposed the Heterogenous Market Hypothesis. The hypothesis is characterized by different market actors having heterogenous time horizons, trading frequences and reactions to news, and they are even likely to settle at different prices. Corsi (2009) simplified these characterics to three types of traders, each generating a different volatility component; short-term traders with daily or higher trading frequency, medium-term traders who rebalance weekly and long-term participants with horizons of one-month or more. Motivated by this hypothesis and empiricial findings revealing volatility cascades, he proposed a volatility cascade time series model with three heterogeneous components:

$$RV_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1d} \tag{3}$$

Where $RV_t^{(d)}$ is the daily realized volatility measure,

$$RV_t^{(w)} = \frac{1}{5} \sum_{i=t}^{t-4} RV_i^{(d)} \tag{4}$$

and

$$RV_t^{(m)} = \frac{1}{22} \sum_{i=t}^{t-21} RV_i^{(d)} \tag{5}$$

The model was estimated by standard OLS regression, using Newey-West standard errors to account for the possibility of serial correlation (Corsi, 2009).

HAR-RV is not formally a long-memory model, but Corsi (2009) showed that simulated autocorrelations of daily realized volatilities, over a 600-year period, yielded the long memory property that was desired. In forecasts of one day, one week and two weeks ahead, the HAR-RV model performed well in-sample and out-of-sample. It outperformed the simple AR(1) and AR(3) regressions, and was comparable to ARFIMA (Corsi, 2009). These initial performance results have later been supported by the work of Ma et al. (2014). In Model Confidence Set tests, comparing the out-of-sample forecast performance of several high-frequency volatility models, including ARFIMA-RV, HAR-RV outperformed the other models.

2.1.2 Microstructure noise

An issue several researchers have faced and discussed, including Corsi (2009) and Andersen et al. (2001), is the presence of microstructure noise in the realized volatility measure. In a frictionless and continuous financial market, ultimately, the highest possible data frequency would give the most precise estimates and forecasting ability. However, the presence of microstructure effects such as nonsynchronous trading effects, bid-ask spreads, discrete price observations and intraday periodic volatility patterns make this impossible (Andersen & Bollerslev, 1998). The effect of noise on the estimator of realized volatility depends on the sampling scheme, and at what frequency it is sampled.

Corsi (2009) employs the Zhang et al. (2005) estimator, known as the Two-Scales Estimator (TSE). This is most appropriate for higher frequency data, such as tick-by-tick data (1 s sampling), i.e. a sample size of 23 400 observations per day. TSE takes advantage of the fact that e.g. 5-minute returns starting at 09:30 can be calculated using intervals 09:30-09:35, 09:35-09:40 and so on, or 09:31-09:36, 09:36-09:41 and so on. Hence, one can create K subsets with 5-minute log returns starting at 09:30 ($k = 1$), 09:31 ($k = 2$), …, 09:34 ($k = 5 = K$), calculate the realized variance for each subset and compute the arithmetic mean. This estimate is then adjusted by a noise factor, which exploits that the full sample realized variance measure is an estimator for the variance of microstructure noise, assuming that noise observations are independent and identically distributed. See Table 9-1 in Appendix 1 for visualization of the example. TSE is specified in equation (6):

$$TSE = RVar_t^{Zhang\ et\ al.} = \frac{1}{K}\sum_{1}^{K} RVar_t^k - \frac{\bar{n}_t}{n_t} RVar_t^{(all)} \tag{6}$$

Where n_t is the number of observations in the full sample, $\bar{n}_t = \frac{n_t - K + 1}{K}$ and $RVar_t = \sum_{i=0}^{M-1} r_{t,i}^2$.

A refinement, see equation (7), is suggested by Aït-Sahalia et al. (2011) for sample sizes smaller than those from tick-by-tick data, and this estimator is unbiased to a higher degree than equation (6):

$$RVar_t^{\text{Aït−Sahalia et al.}} = \left(1 - \frac{\bar{n}_t}{n_t}\right)^{-1} RVar_t^{Zhang\ et\ al.} \tag{7}$$

A popular way of dealing with microstructure noise is to use sparse sampling, i.e. sampling at arbitrary lower frequencies, like 5, 10, 15 or 30 minutes. This is used by Andersen et al. (2001, 2003). It was shown in Andersen and Bollerslev (1998), that 5-minute frequency is high enough to reduce measurement error, but low enough to avoid major microstructure noise bias.

2.2 Data

The measure of daily realized volatility is the fundamental measure from which all other volatility variables in our data set are calculated, and is the dependent variable in our analysis. The raw stock price data is collected in two parts, due to availability issues. The first part contains 5-minute frequency price data from December 3rd 2019 to January 5th 2020. The second part contains 1-minute frequency price data in the interval January 6th to April 3rd 2020. Although there are observations in the raw stock price data outside of trading hours, these observations are infrequent and the transaction volume is relatively low. Therefore, any observations outside of trading hours are discarded. So, for the 5-minute and 1-minute frequencies, the full trading day market data contains 79 and 391 observations of the continuously compounded return for each stock, respectively. All market data is obtained from Thomson Reuters Eikon.

As shown in Table 2-1, there are some missing price observations in the raw, 1-minute frequency data. From a total of 541 926 price realizations, 1 610 observations are missing. 1 350 are missing due to market wide trading halts that affected all 22 stocks. 237 are missing first trading-minute observations, where the previous price observation is no longer than 30 minutes earlier. The remaining 23, mainly consist of single observations and at most four consecutive observations.

Table 2-1 Missing observations in 1-minute frequency data

Description of missing observations	Quantity	% of total observations
Missing due to market wide trading halts	1 350	0,249 %
First trading minute	237	0,044 %
Other	23	0,004 %
Total missing within trading hours	1 610	0,297 %

To ensure that our formulas are applied consistently throughout the data, we assume that any missing observations arise from one of two things: either no trading activity or no change in price. Therefore, we simply use the previous price observation in any calculations involving a missing observation.

2.2.1 Two-Scales Estimator

Using TSE, we calculate the daily observation of realized volatility from the 1-minute frequency data. We make the assumption that 1-minute frequency is sufficiently frequent to ensure that the full sample realized variance is an estimator of the variance of microstructure noise. We also use the small sample refinement suggested by Aït-Sahalia et al. (2011) to reduce any bias arising from this assumption.

We create $K = 10$ subsets with 10-minute log-return intervals. The first subset, $k = 1$, starts at 09:30 when trading hours start. The first log return is calculated at 09:40 and the final log return at 16:00, when trading hours close. The remaining 9 subsets, $k = [\,2\,,10\,]$, start at 09:30 $+ (k - 1)\ minutes$. The first log return is calculated at 09:40 $+ (k - 1)$ and the final log return at 15:50 $+ (k - 1)$. Hence, subset 10 starts at 09:39 and stops at 15:59, which means that it fails to account for the full trading day. To avoid underestimating the daily volatility in subset [2 , 10], we also calculate the log return from 09:30 to $09{:}30 + (k - 1)$ and from $15{:}50 + (k - 1)$ to 16:00. This way we ensure that the volatility measure from all subsets are calculated over the same 6,5-hour time period.

We create the remaining two variables, $RV_t^{(w)}$ and $RV_t^{(m)}$, needed to estimate the HAR-RV model using equation (4) and equation (5). To create $RV_t^{(w)}$ and $RV_t^{(m)}$ for the first observation in the data set, January 6th, observations of daily realized volatility dating back to December 27th and December 3rd is needed, respectively. Therefore, we make use of the 5-minute frequency data. The daily realized volatility from December 3rd to January 5th are calculated using equation (2). Although the presence of microstructure noise and measurement error are likely to be higher in these estimates, compared to estimates using TSE, we propose that it is appropriate to use these estimates to create $RV_t^{(w)}$ and $RV_t^{(m)}$. We believe the alternative, reducing the sample by 22 trading days, to yield a less precise model estimation. Assuming that microstructure noise and measurement errors are independent and identically distributed, with mean equal to zero, the volatility estimates averaged over 5 and 22 days should be sufficiently accurate.

2.2.2 Summary statistics

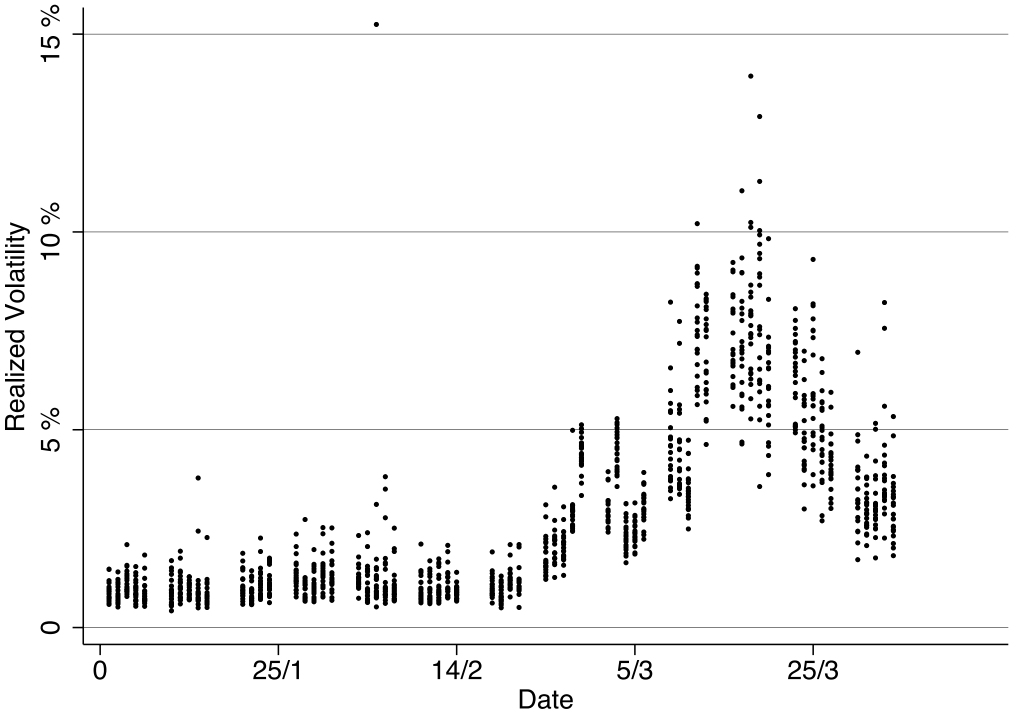
In Table 2-2, we present summary statistics for our realized volatility variables. We see that the means for all three variables are relatively high, considering that the sample consists of 22 companies from S&P 100. As seen in Figure 2-1, there is a large spike in the timeline due to the effect of Covid-19 on financial markets, which is pulling the means upward. However, the median of Rvola is 1,7 %, which is considerably lower.

Table 2-2 Summary statistics for volatility variables

Variable name	Description	Observations	Mean	Std. Dev	5%	50%	95%
Rvola	Realized volatility	1 386	2,7 %	2,3 %	0,7 %	1,7 %	7,6 %
Rvolawk	$RV_t^{(w)}$	1 386	2,6 %	2,2 %	0,8 %	1,4 %	7,0 %
Rvolamt	$RV_t^{(m)}$	1 386	1,9 %	1,4 %	0,8 %	1,2 %	5,0 %

Figure 2-1 shows the daily observations of realized volatility for each stock over time. The effect of Covid-19, from February 24th, creates a large spike in the latter part of our sample. Controlling for time fixed effects, or market volatility, is therefore important in the analysis of this data set. Removing cross-sectional means and running a Levin-Lin-Chu test (Levin et al., 2002), we can reject that realized volatility exhibits a unit root on at least a 0,1 % level, see Table 9-2 in Appendix 2. We can also see a large outlier on February 5th. This is Biogen Inc., with a realized volatility estimate of 15,24 %. Biogen opened at $285,63 and closed at $332,87 per share.

Figure 2-1 Realized volatility over time period



3 Twitter

In this section, we present arguments in favor of Twitter as a predictor of financial data, as well as detailed information about our suggested attention indicators. Further, we outline the data collection process, and present all Twitter variables used to perform analysis.

3.1 Theory

Several arguments can be made in favor of Twitter as a data source. Twitter is the most popular microblogging site, hosting many influential people and organizations with broad followings. In line with similar sites, it allows users to react to events and interact with other users through a short message system. Compared to traditional mediums, e.g. internet message boards, Twitter allows for more real-time conversation that responds quickly to news, making intricate analysis between social media and events in financial markets more interesting. Further, the messages have an upper limit of 140 to 280 characters, depending on region, and the brevity of the messages require users to be more concise, as well as making data processing more manageable. Moreover, its free API permits access to an extensive library of messages that are easy to extract, as opposed to traditional research instruments that are costly to develop. Lastly, filtering messages using searchable operators and keywords, make it easy to navigate and find data. Our study takes advantage of these features and utilizes cashtag and company name mentions to create attention indicators.

3.1.1 Cashtags

Cashtag is a searchable operator on Twitter that comprises the dollar symbol followed by the company ticker symbol, e.g. $MSFT, which relates to specific company stock. The selection and filtering of messages using cashtags permits analysis of individual stocks, and could also mitigate noise in the data set by excluding messages that are less related to markets and market participants. An introductory study by Hentschel and Alonso (2014) looked at how widespread the application of cashtags were and how they conveyed financial information on Twitter. Their findings suggested that messages containing cashtags were associated with financial activity, as some stock prices were found to correlate with spikes in cashtag message volume. Moreover, cashtags have been crucial to understanding the impact of Twitter sentiment on markets. Studies from Sprenger et al. (2014) and Oliveira et al. (2017) both found messages containing cashtags to be valuable for predicting market variables.

3.1.2 Company names

The selection of messages using company names, e.g. Cisco Systems, pertains to any information regarding a company. In contrast to cashtags, variables relying on these messages may be prone to larger measurement errors, as people refer to companies in multiple different ways. However, they may still contain valuable information that is overlooked if only cashtags are applied. In order to capture the most relevant messages, and increase precision, endings for company names that have Corp, Inc and Co in them are omitted, as they are likely excluded from conversations on Twitter.

3.1.3 On the issue of spam messages

A common issue with Twitter messages is the presence of spam. Hentschel and Alonso (2014) stress the importance of separating spam messages from legitimate user messages, and further show that some companies are prone to spam. An inherent property of Twitter's free search API, is that it emphasizes relevance over completeness. Although Twitter fails to provide an official elaboration on this, a study conducted by Thelwall (2015) indicate that most messages of importance are returned, even if the samples are incomplete. Further, he argues that excluded messages are unlikely to be problematic for research purposes, as most of the removed messages from his sample were either duplicates or had spam characteristics. We recognize that spam presents an issue for any study that relies on Twitter data, however, our decision to use the free search API is a conscious effort to reduce this problem, due to its inherent filtering properties.

3.2 Data

Twitter's free search API searches against a sampling of published messages in the last seven days. Our searches were executed using rtweet, a community developed R application listed on Twitter's developer page. rtweet enables users to access the API endpoints, using the programming language of R instead of HTTP and JSON. The standard search API has a rate limit of 18 000 messages every 15 min and per request. Using rtweet, we were able to circumvent this limit by enabling the option to retry on rate limit. This ensures that any halted searches continue after the rate limit resets.

For each of the 22 stocks in the data set, searches were conducted on a weekly basis. We ran searches for messages containing the company specific cashtag, collecting messages from the last 7 days. The same searches were done for messages containing the company name. In total, the data set includes Twitter messages spanning from January 5th to April 2nd 2020.

We create three variables from each of the two collections of Twitter messages by calculating the message volume for each stock in specific time intervals, see Table 3-1. First, we calculate the daily volume from 00:00 to 23:59, in variables we label Ct for messages containing cashtags, and Nm for company name. Although volume is low during the night, Ct and Nm forgo any messages in the hours leading up to market openings. Therefore, we create two 24-hour variables that accumulate messages from 09:30 on day $t-1$, until just before the market opens at 09:29 on day t. We label these variables Ct-24h and Nm-24h. However, relevant market information on Twitter posted during trading hours might already be reflected in volatility the same day. Thus, we create two variables that capture all messages from market closings at 16:00 on day $t-1$, until just before the market opens at 09:29 on day t. We label these variables Ct-17,5h and Nm-17,5h.

3.2.1 Summary statistics

Table 3-1 presents summary statistics for Twitter variables.

Table 3-1 Summary statistics for Twitter variables

Variable name	Description	Observations*	Mean	Std. Dev
Cashtag				
Ct	Full day message volume: 00:00 - 23:59	1 892	111,4	148,9
Ct-24h	Message volume 24 hours prior to trade opening: 09:30(t - 1) - 09:29	1 936	114,0	147,5
Ct-17,5h	Message volume since trading closed yesterday: 16:00(t - 1) - 09:29	1 936	64,2	88,3
Company name				
Nm	Full day message volume: 00:00 - 23:59	1 890	604,4	1 587,3
Nm-24h	Message volume 24 hours prior to trade opening: 09:30(t - 1) - 09:29	1 932	606,4	1 574,4
Nm-17,5h	Message volume since trading closed yesterday: 16:00(t - 1) - 09:29	1 933	394,4	1 139,4

* Differing number of observations are due to missing messages from February 27th 12:00 to February 28th 02:00 because of scheduling error and for keyword search "Bank of America" there are no messages prior to January 8th 16:00 due to an unknown error.

As we see in Table 3-1, there is a big difference in mean volume and standard deviation for cashtag and company name searches. Furthermore, an increase in company name characters are associated with lower message volumes, as evidenced by a correlation coefficient of -0,28 between number of characters and mean volume for each company name search. This indicates that longer names are more likely to be abbreviated when microblogging.

3.2.2 Non-business days

Autoregressive models of volatility rely on lags of volatility, and specifically the first lag in the HAR-RV model. When estimating volatility following non-business days, we use the volatility measure from the previous trading day. However, this is not a suitable solution for Twitter, as information continues to spread throughout weekends and non-business days. We propose two candidate solutions: aggregating message volume over non-business days, or using the previous day message volume regardless of whether trading occurred on this day. This produces Table 3-2, with variables that aggregate volume for non-business days and forwards them to the next business day.

Table 3-2 Summary statistics for Twitter variables aggregating non-business days

Variable name	Description	Observations*	Mean	Std. Dev
Cashtag				
Ct-sumred	cashtag variable + sumred	1 298	130,8	166,4
Ct-24h-sumred	Ct-24h + sumred	1 342	160,1	205,3
Ct-17,5h-sumred	Ct-17,5h + sumred	1 342	95,4	136,6
Company name				
Nm-sumred	name + sumred	1 297	693,0	1 821,2
Nm-24h-sumred	Nm-24h + sumred	1 339	855,5	2 055,6
Nm-17,5h-sumred	Nm-17,5h + sumred	1 340	585,1	1 544,4

* The reduction in observations from Table 3-1 is due to the removal of non-business day observations

4 The panel

To study the empirical relationship between stock market data and Twitter messaging data, and the forecasting ability of Twitter message volume for the next day volatility of stock returns, we have created a panel consisting of a random sample of 22 stocks from the 100 largest publicly listed corporations in the US. The population in mind for this paper is the corporations listed on the S&P 100. However, this study could be externally valid for similar populations where Twitter is widely used. We narrow down to a limited population because Twitter's standard search API prohibits historical searches beyond the last seven days. As such, there is a constraint on the amount of observations per company in this paper, which could threaten the significance of our estimators. It is also likely that our data contains some form of noise or measurement error. Therefore, we chose a population which has some of the most liquid stocks in the world, and where Twitter is widely used. Usage in the US accounted for 20% of "Monetizable Daily Active Usage" on Twitter in 2019 (Twitter Inc, 2019b, p. 1).

4.1 Panel summary statistics

Table 4-1 presents summary statistics for each of the 22 companies regarding central tendencies of volatility, and the distribution of message volume for our selected search words and cashtags. Comparing the means of realized volatility and Twitter message volume for each company, the figures suggest no obvious relationship between the means. Kinder Morgan has the highest mean volatility, yet mean Ct and mean Nm is in the bottom 15th and 25th percentile, respectively. Meanwhile, Costco Wholesale has the lowest mean volatility, a mean Ct in the top 25th percentile and mean Nm in the bottom 10th percentile. Although this needs to be considered more carefully, the differences in means indicate that changes in the variables might be more interesting than absolute levels. Also, a qualitative assessment in choice of search words might have produced different means, e.g. "Cisco" instead of "Cisco Systems", "Costco" instead of "Costco Wholesale".

Table 4-1 Panel summary statistics

Company name*	Cashtag	Mean Ct	Mean Nm	Mean Rvola	Median Rvola
American Express	$AXP	58	1143	2,9 %	1,6 %
Bank of America	$BAC	215	1774	2,8 %	1,3 %
Biogen	$BIIB	102	683	3,1 %	2,1 %
Bristol-Myers Squibb	$BMY	86	75	2,4 %	1,6 %
Cisco Systems	$CSCO	102	77	2,6 %	1,4 %
Costco Wholesale	$COST	176	37	2,0 %	1,3 %
Danaher	$DHR	25	63	2,5 %	1,6 %
Emerson Electric	$EMR	14	49	3,0 %	1,7 %
General Dynamics	$GD	29	128	2,6 %	1,4 %
Gilead Sciences	$GILD	354	393	2,8 %	2,4 %
Johnson & Johnson	$JNJ	176	428	2,2 %	1,2 %
JPMorgan Chase	$JPM	276	580	2,7 %	1,5 %
Kinder Morgan	$KMI	28	65	3,2 %	1,2 %
Eli Lilly	$LLY	65	388	2,5 %	1,4 %
Lockheed Martin	$LMT	87	867	2,6 %	1,6 %
Medtronic	$MDT	45	1195	2,5 %	1,4 %
Qualcomm	$QCOM	134	1353	3,0 %	2,2 %
Thermo Fisher Scientific	$TMO	28	364	2,6 %	2,0 %
Union Pacific	$UNP	29	88	2,8 %	1,7 %
Walgreens Boots Alliance	$WBA	51	17	3,0 %	1,7 %
Wells Fargo	$WFC	136	3133	2,8 %	1,4 %
Exxon Mobil	$XOM	235	422	3,0 %	1,6 %

* Company name here refers to the search words used, not the actual name of the company.

In the analysis and results, we use scaled variables in the regression models to ease readability of coefficients. All volatility variables will be scaled by a factor of 100, and now represent percentage points. All Twitter variables will be scaled by a factor of 1/100, and now represent message volume measured in hundreds.

4.2 Specifying the HAR-RV model with panel data

The HAR-RV model, as proposed by Corsi (2009), used a time series framework where the model estimated the specific properties of a single financial instrument. The data set employed in this analysis is a panel consisting of 22 stocks, which demands different considerations than in a single entity model.

First, we take a look at the HAR-RV model in a panel data setup with entity fixed effects:

$$RV_{i,t+1d}^{(d)} = \beta_0^{(d)} + \beta^{(d)} RV_{i,t}^{(d)} + \beta^{(w)} RV_{i,t}^{(w)} + \beta^{(m)} RV_{i,t}^{(m)} + \alpha_i + u_{i,t+1d} \tag{8}$$

Where i represents each entity, α_i is a collection of dummy variables representing the entity fixed effects and $u_{i,t+1d}$ is the error term. In this context, the entity fixed effects control for inherent differences in volatility across stocks that remain constant over time. Together with the constant $\beta_0^{(d)}$, α_i allows the long term volatility estimate to vary across stocks.

In Table 4-2, we present the OLS regression results from estimating the HAR-RV model on a panel of 22 entities, using Newey-West standard errors of order 5. Note that the dummy for each entity is not included in the table, instead the joint F-statistic is provided. L1.Rvola represents the first lag of Rvola. The model overall appears to be a good fit, but the monthly volatility variable is estimated to have a negative coefficient. This is counterintuitive. Due to the well-known long memory properties of volatility, we would expect all lagged components of volatility to have positive coefficients. Furthermore, the volatility measure is an absolute measure of risk. Hence, we would expect the model to always predict positive values. Upon closer inspection, the model also suffers from multicollinearity issues. Furthermore, the entity fixed effects, α_i, are not jointly significant with an F-stat equal to 0,17.

Table 4-2 Panel HAR-RV model regression

Rvola	Coefficients
L1.Rvola	0,427***
	(7,57)
Rvolawk	0,639***
	(9,14)
Rvolamt	-0,265***
	(-5,68)
α_i	-
	(0,17)
Constant	0,417**
	(2,31)
Observations	1342
F-test	114,1
p-value	9,00e-301

t-statistics in parentheses and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$,

Joint F-statistic for α_i in brackets.

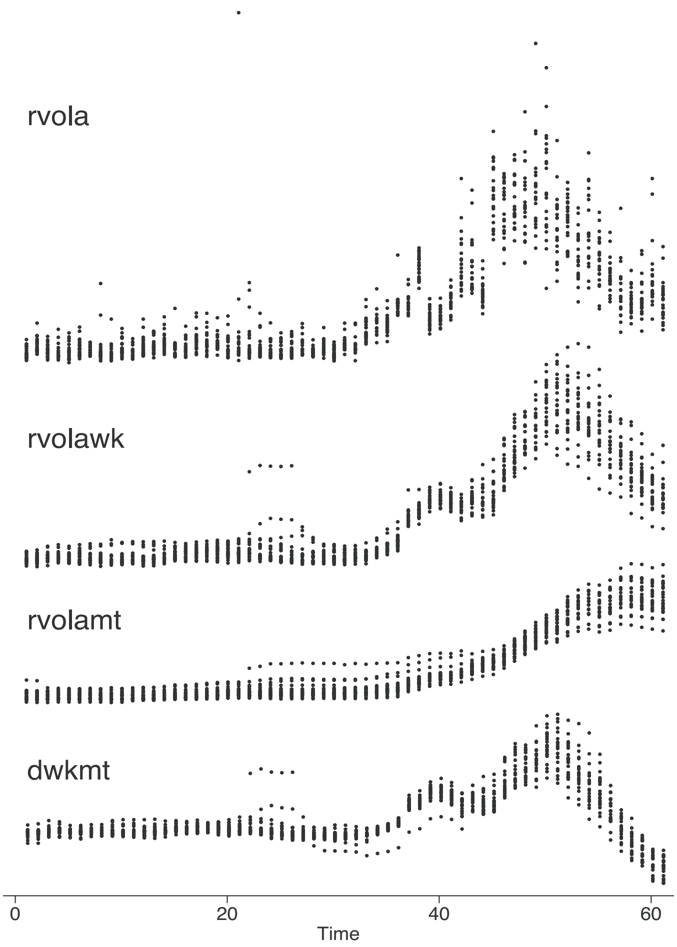
We propose a transformation of the variables and create a new variable, Dwkmt. The new variable is the difference between the weekly and the monthly volatility components:

$$Dwkmt = RV_t^{(w)} - RV_t^{(m)} \tag{9}$$

The economic intuition behind this variable is that recent measures of volatility are a better predictor of volatility, so in a case where the difference between the average over the past week and the past month increases, the model should suggest an increase in realized volatility for the next day, and vice versa.

Figure 4-1 is a visual comparison of Rvola, Rvolawk, Rvolamt and Dwkmt along the timeline of our sample. Rvolawk appears to be highly dependent of Rvola, but it responds slightly slower to large spikes in volatility. Rvolamt is even less responsive than Rvolawk. It appears to lag behind Rvola and Rvolawk, and is more important in setting the level of Rvola over a longer time period. Dwkmt appears to be highly dependent of Rvolawk and responds similarly to spikes, but adjusts further down towards the end of a spike. Rvolamt and Dwkmt appear to respond conversely at the end of the spike, and the combination of these two appear to be a good fit. We compare these findings to the correlation matrix in Table 9-3, in Appendix 3, which confirms that Dwkmt is highly correlated with Rvolawk, with a coefficient equal to 0,77. Dwkmt is also correlated to Rvolamt, but less so, with a coefficient equal to 0,35.

Figure 4-1 Studying Dwkmt



Note that there is no vertical axis presented in this figure, only the co-movement between the four scattered clouds are of interest.

We propose employing Dwkmt as a replacement for Rvolawk. With the inclusion of Dwkmt, we still make use of all the data and components of the HAR-RV model, but we estimate the weekly component jointly with the monthly component.

These alterations produce our baseline HAR-RV model, that will be applied and tested against when forecasting in hypothesis 3. The baseline HAR-RV model is specified in equation (10):

$$
\begin{aligned}
RV_{i,t+1d}^{(d)} = \beta_0^{(d)} + \beta^{(d)}\left(RV_{i,t}^{(d)} - \overline{RV}_i^{(d)}\right) + \beta^{(m)}\left(RV_{i,t}^{(m)} - \overline{RV}_i^{(m)}\right) \\
+ \beta^{(w-m)}\left(RV_{i,t}^{(w)} - RV_{i,t}^{(m)}\right) + \alpha_i + u_{i,t+1d}
\end{aligned}
\tag{10}
$$

Where $\overline{RV}_i^{(d)}$ is the average L1.Rvola for entity i, and $\overline{RV}_i^{(m)}$ is the average Rvolamt for entity i.

Table 4-3 presents our baseline HAR-RV model. Findings reveal that our new model specification is also a good fit, and that the proposed variable transformations reduce issues pertaining to multicollinearity and negative regressors.

Table 4-3 Baseline HAR-RV model

Rvola	Coefficients
L1.Rvolac	0,427***
	(7,57)
Rvolamtc	0,373***
	(7,74)
Dwkmt	0,639***
	(9,14)
α_i	-***
	(3,58)
Constant	2,421***
	(12,75)
Observations	1342
F-test	114,1
Prob > F	0,000

t-statistics in parentheses and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$,

Joint F-statistic for α_i in brackets.

Further, we find that Mean Variance Inflation Factor (VIF) for the three volatility variables has been reduced from 8,10 to 4,34. The new variable, Dwkmt, is significant and has a positive coefficient in accordance with the economic intuition explained earlier. Moreover, we find that centering L1.Rvola and Rvolamt, by subtracting their respective panel means, results in jointly significant entity fixed effects at a 0,1% level. Note that we add "c" to the variable name when centering.

4.3 On the issue of heteroskedasticity, serial and cross-panel correlation

It is well-known that volatility modeling encounters issues of heteroskedasticity and serially correlated residuals. Volatility has been shown to be both persistent and mean-reverting (Engle & Patton, 2007). These are features that Corsi (2009) seek to model with HAR-RV. Still, a Cumby-Huizinga test (Cumby & Huizinga, 1992) reveals that residuals from the baseline HAR-RV model, estimated with OLS, are serially correlated, see Table 9-4 in Appendix 4. All results from tests in subsection 4.3 can be found in Appendix 4. Graphical examination and a Breusch-Pagan test (Breusch & Pagan, 1979) reveal that residuals are heteroskedastic, see Figure 9-1 and Table 9-5. Therefore, standard errors in this paper are calculated using the Newey-West heteroskedasticity and autocorrelation consistent covariance matrix of order 5 (Newey & West, 1987). This is consistent with how standard errors are calculated in Corsi (2009) and Patton and Sheppard (2015).

Applying the HAR-RV model in a panel data setup induces a third issue, namely cross-panel correlation. The daily measures of volatility in a panel of 22 stocks from the S&P 100 are likely to be highly dependent in the cross-section, as they all are affected by the same macroeconomic environment. A general cross section dependence test (Pesaran, 2004) (Pesaran, 2015) reveal that the residuals are correlated across panels, see Table 9-6. However, controlling for time fixed effects drastically reduces the correlation, but it does not remove dependence completely, see Table 9-7. So, in order to increase the validity of this paper's findings, we conduct tests using bootstrapped standard errors, and sample from both entity and time clusters. See Table 9-8 and Table 9-9 for bootstrap results for the baseline HAR-RV model.

5 Empirical approach

In this section, we outline the empirical approach used to test our hypotheses.

5.1 Approach for H1 and H2

Equation (11) to (13) specify the regression models applied to test H1 and H2, for each Twitter variable. We propose an approach in which we gradually introduce control variables to study the dynamics of the relationships. The variable $[Twitter]$ is simply a placeholder for each Twitter variable.

In the Fe model (11), we test the raw relationship between Twitter message volume and Rvola, only controlling for entity fixed effects:

$$Rvola_{i,t} = \beta_0 + \beta_1[Twitter]_{i,t-1d} + \alpha_i + u_{i,t} \tag{11}$$

The HAR + Fe model (12) tests the relationship when we also control for variation explained by our baseline HAR-RV model:

$$\begin{aligned} Rvola_{i,t} = \beta_0 &+ \beta_1[Twitter]_{i,t-1d} + \beta_2 L1.Rvolac_{i,t-1d} \\ &+ \beta_3 Rvolamtc_{i,t-1d} + \beta_4 Dwkmt_{i,t-1d} + \alpha_i + u_{i,t} \end{aligned} \tag{12}$$

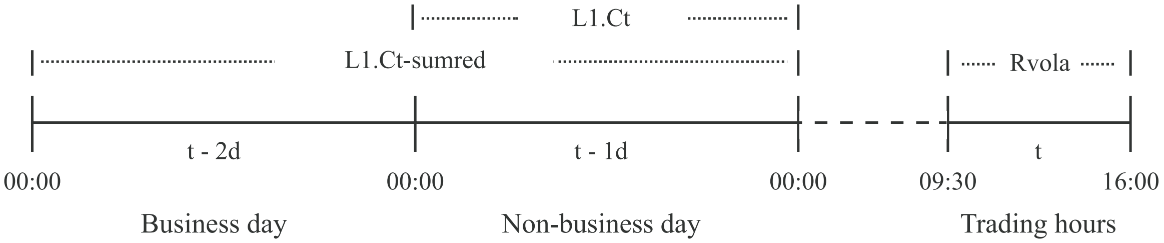
In the HAR + Fe + Te model (13), we also add time fixed effects, controlling for omitted variables that vary over time but are constant across entities, e.g. market volatility:

$$\begin{aligned} Rvola_{i,t} = \beta_0 &+ \beta_1[Twitter]_{i,t-1d} + \beta_2 L1.Rvolac_{i,t-1d} \\ &+ \beta_3 Rvolamtc_{i,t-1d} + \beta_4 Dwkmt_{i,t-1d} + \alpha_i + \lambda_t + u_{i,t} \end{aligned} \tag{13}$$

To better visualize how Twitter variables relate to Rvola in time, we illustrate the relationship with timelines in Figure 5-1, Figure 5-2 and Figure 5-3. The timelines are illustrated with cashtag variables. However, name variables hold the same relationship to Rvola in time. Therefore, the timelines and reasoning that follow also apply to the equivalent name variables, e.g. L1.Nm is analogous to L1.Ct.

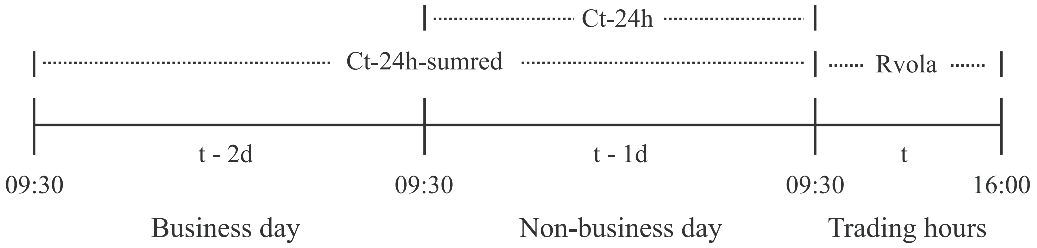
As we see in Figure 5-1, L1.Ct contains messages that are accumulated between 00:00 and 23:59 the day previous to the dependent variable Rvola, regardless of whether $t - 1d$ is a business day or not. L1.Ct-sumred also accumulates messages over non-business days.

Figure 5-1 Timeline L1.Ct and L1.Ct-sumred



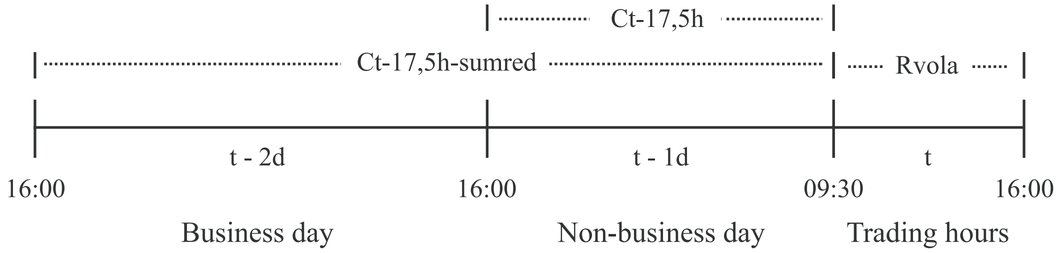
As we see in Figure 5-2, Ct-24h contains messages from the last 24 hours before trading starts on day t, regardless of whether $t - 1d$ is a business day or not. Ct-24h-sumred also accumulates messages over non-business days.

Figure 5-2 Timeline Ct-24h and Ct-24h-sumred



As we see in Figure 5-3, Ct-17,5h contains messages from the last 17,5 hours before trading starts on day t, regardless of whether $t - 1d$ is a business day or not. Ct-17,5h-sumred also accumulates messages back to 16:00 on the last business day.

Figure 5-3 Timeline Ct-17,5h and Ct-17,5h-sumred



5.2 Approach for H3

To test H3, we conduct both in-sample and pseudo out-of-sample forecasts. We base our choice of forecasting models on the performance of Twitter variables in H1 and H2. The best performing cashtag and name variable will be used in conjunction with our baseline HAR-RV model.

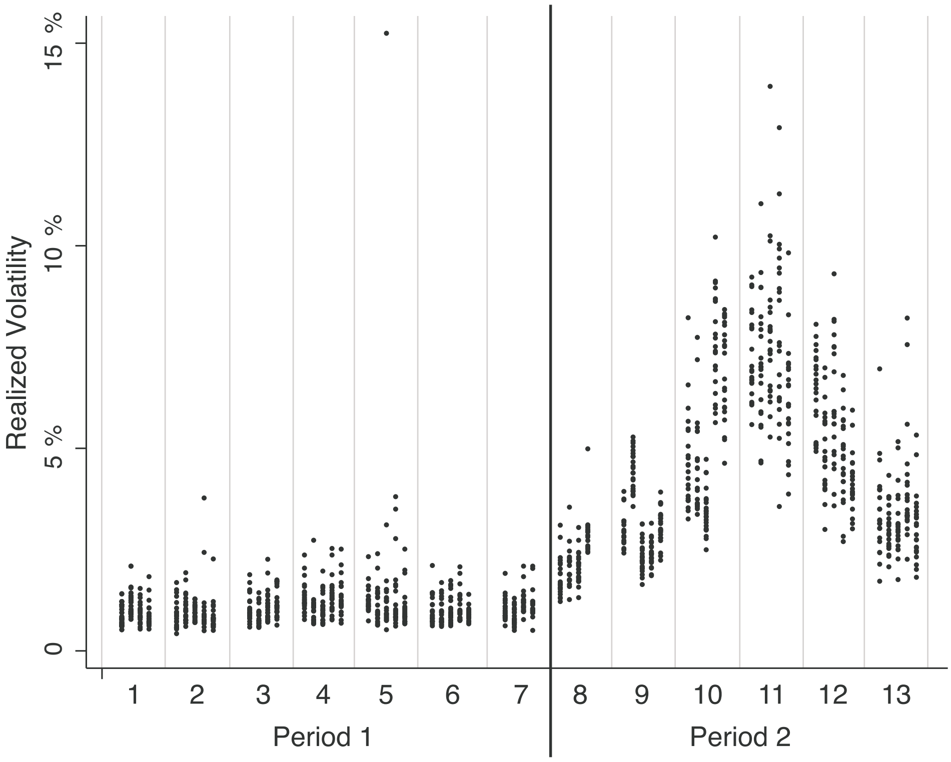
The combination of a short sampling period, and the effect of Covid-19, presents a challenge in how we approach forecasting. Specifically, the magnitude of the Covid-19 related spike makes forecasts sensitive to subsample selection. Running a Chow test (Chow, 1960), we find a break in the time series on February 24th, with an F-stat equal to 40,52. Forecasting with time fixed effects is infeasible, therefore, we propose a solution in which we forecast using a dummy variable for the observed break on February 24th that interacts with each independent variable. This should ensure that forecasting performance is consistent throughout the sample.

The full sample is divided into 13 subsamples; one for each calendar week. We estimate the models by creating a rolling window. For each iteration of the forecasting process, the rolling window excludes subsample i from the estimation process, then estimates the model, and finally, forecasts into the excluded subsample i. Thus, we can test each model's performance on 13 candidate scenarios of realized volatility in $t+1$ to $t+5$.

It is important to note that the rolling window does not fully exclude observations of Rvola from the estimation process. To estimate models containing HAR-RV consistently, observations of Rvola in the excluded subsample are included in Rvolawk and Rvolamt for subsequent estimates. For instance, if observations of Rvola at time $t-1$ are excluded from Rvolawk at time t, then we are not properly estimating the HAR-RV model at time $t+1$. Therefore, the observations of Rvola in the excluded subsample must be included in subsequent Rvolawk and Rvolamt calculations. Although this subsampling process gives rise to concerns about validity, due to issues regarding our sampling period, we believe this procedure will produce more robust forecasts than conventional methods of subsampling for this data set.

The changing macroeconomic environment throughout our sample enables testing of Twitter variables' forecasting performance under steady market conditions, and during a period of turmoil. We group subsamples 1 to 7 into Period 1, and subsamples 8 to 13 into Period 2. The break date is the first trading day in subsample 8. In Figure 5-4, we visualize the two forecasting periods, illustrating the difference in volatility levels and spread between subsamples. Now, Period 1 is located before Covid-19 affects the US financial markets, while Period 2 is amid the Covid-19 crisis.

Figure 5-4 Forecasting periods and subsamples



To study the forecasting performance of our models, we adopt two loss functions that rely on standardized forecasting errors, see equation (14) and (15), and one that is widely accepted in the literature, see equation (16).

Considering the work of Patton (2011) on robust loss functions for realized volatility, we employ the QLIKE measure, which is specified in equation (14):

$$QLIKE = \frac{1}{N * T} \sum_{i}^{N} \sum_{t}^{T-1} \frac{RV_{i,t+1}^{2}}{f_{i,t+1}^{2}} - \ln\left(\frac{RV_{i,t+1}^{2}}{f_{i,t+1}^{2}}\right) - 1 \qquad (14)$$

Where, $f_{i,t+1}$ is the forecasted value of realized volatility for stock i at time $t + 1$.

Further, we adopt mean absolute percentage error (MAPE) because it is less sensitive to outliers. MAPE is specified in equation (15):

$$MAPE = \frac{1}{N * T} \sum_{i}^{N} \sum_{t}^{T-1} \frac{\left| RV_{i,t+1} - f_{i,t+1} \right|}{f_{i,t+1}} \tag{15}$$

Finally, we employ root mean squared error (RMSE), which is specified in equation (16):

$$RMSE = \sqrt{\frac{1}{N * T} \sum_{i}^{N} \sum_{t}^{T-1} \left(RV_{i,t+1} - f_{i,t+1} \right)^{2}} \tag{16}$$

All three measures will be averaged for Period 1 and Period 2, respectively, to study differences in performance for the two periods.

Putting performance measures in a statistical context, we test differences in predictive accuracy using the Diebold-Mariano test, which tests whether differences in an arbitrary loss function from two competing forecasts are statistically significant (Diebold & Mariano, 1995):

$$DM_{i,j} = \frac{\bar{d}_{i,j}}{\hat{\sigma}_{\bar{d}_{i,j}}} \xrightarrow{d} N(0,1) \tag{17}$$

Where $\bar{d}_{i,j}$ is the average difference in a loss function between model i and j, and $\hat{\sigma}_{\bar{d}_{i,j}}$ is an estimator for the standard error of $\bar{d}_{i,j}$. We estimate the DM-statistic using bootstrap estimation with 5000 repetitions, similarly to the estimation used in the Model Confidence Set test (Hansen et al., 2011).

6 Results and discussion

In this section, we present the results and corresponding discussion for hypothesis 1, 2 and 3.

6.1 Hypothesis 1

H1: Changes in company attention on Twitter, measured by message volume related to company stock accumulated prior to trading hours, is associated with changes in volatility.

To test H1, we propose that cashtag variables are a suitable approximation of all messages related to company stock.

6.1.1 Results

In Table 6-1, we present results from regressions using the six cashtag variables as explanatory variables, in accordance with the approach explained in section 5.1. Note that the regression results only include the coefficients for cashtag variables, as these are the coefficients of interest.

As is shown for Fe models, coefficients for cashtag variables are all positive values and significant on a 1 % level. However, none of the variables hold enough explanatory power in conjunction with entity fixed effects to produce significant model F-tests.

In the HAR + Fe models, L1.Ct and L1.Ct-sumred are not significant. Ct-24h and Ct-24h-sumred are both significant on a 5 % level and an increase of 100 messages is associated with approximately a 0,06 %-points increase in Rvola. Ct-17,5h and Ct-17,5h-sumred are significant on a 1 % level and an increase of 100 messages is associated with an increase in Rvola of approximately 0,14 and 0,09 %-points, respectively.

HAR + Fe + Te models are consistently the best models in terms of F-stat, and the coefficients for L1.Ct, Ct-24h and Ct-17,5h all have higher t-stats than their "sumred" counterparts. L1.Ct is significant on a 5 % level, while both Ct-24h and Ct-17,5h are significant on a 1 % level. On average, controlling for time fixed effects induces a 17 % reduction in coefficients for Ct-24h, Ct-24h-sumred, Ct-17,5h and Ct-17,5h-sumred.

Table 6-1 Regression results for all cashtag variables

Base		[Twitter]	Coefficient	t-stat	F-stat
Fe	+	L1.Ct	0,2454***	3,55	0,93
HAR + Fe	+	L1.Ct	0,0223	0,84	109,63
HAR + Fe + Te	+	L1.Ct	0,0372**	2,02	174,57
Fe	+	L1.Ct-sumred	0,2011***	2,96	0,72
HAR + Fe	+	L1.Ct-sumred	0,0268	1,00	109,81
HAR + Fe + Te	+	L1.Ct-sumred	0,0286	1,63	174,24
Fe	+	Ct-24h	0,2598***	3,91	1,07
HAR + Fe	+	Ct-24h	0,0649**	2,21	109,91
HAR + Fe + Te	+	Ct-24h	0,0691***	3,65	178,79
Fe	+	Ct-24h-sumred	0,1695***	3,42	0,84
HAR + Fe	+	Ct-24h-sumred	0,0628**	2,56	110,36
HAR + Fe + Te	+	Ct-24h-sumred	0,0433**	2,52	176,95
Fe	+	Ct-17,5h	0,3237***	3,32	0,82
HAR + Fe	+	Ct-17,5h	0,1453***	3,78	110,55
HAR + Fe + Te	+	Ct-17,5h	0,1265***	4,91	179,27
Fe	+	Ct-17,5h-sumred	0,01648***	2,60	0,60
HAR + Fe	+	Ct-17,5h-sumred	0,0951***	3,08	110,55
HAR + Fe + Te	+	Ct-17,5h-sumred	0,0661***	2,80	177,69

Fe refers to entity fixed effects, Te refers to time fixed effects. F-stat refers to the entire model.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Unsurprisingly, introducing HAR and time fixed effects have a large effect on the model F-stat. The best model, in terms of F-stat and t-stat, is HAR + Fe + Te + Ct-17,5h. To test the robustness of this model, we run two bootstrap estimations with 1000 repetitions. The first bootstrap estimation is done by sampling from 22 entity clusters, i.e. 22 stocks. The observed coefficient for Ct-17,5h is equal to the initial estimation and it has a 95 % confidence interval equal to $[0{,}0558; 0{,}1972]$. The second estimation is done by sampling from 61 time clusters. We again observe the same coefficient for Ct-17,5h, with a 95 % confidence interval equal to $[0{,}0769; 0{,}1760]$. Both estimations observe a coefficient that is statistically significant on at least a 0,1 % level.

Excluding Fe models, models containing Ct-17,5h outperform the models containing Ct-24h, and further, Ct-24h outperform L1.Ct.

6.1.2 Discussion

Results from Table 6-1 indicate that cashtag variables are related to market activity, as nearly all models produce statistically significant coefficients. Our findings show that cashtag variables closer to Rvola in time are better predictors of volatility, as variables further from Rvola gradually lose statistical significance. Ct-17,5h is the variable closest to Rvola in time, and interestingly, the only variable containing messages that have not accumulated during trading hours. This implies that markets quickly absorb Twitter information, and assuming that markets are efficient, we would expect Twitter information accumulated at $t - 2d$ to already be reflected in volatility estimates for $t - 1d$. Further, coefficients decrease slightly when we introduce time fixed effects, which could indicate that cashtag variables explain variation pertaining to the entire market. However, the variables remain statistically significant and keep their explanatory power. Surprisingly, this suggests that messages containing cashtags uniquely explain Rvola for our selection of companies, even if we control for market volatility. Due to Covid-19 affecting our data set, we would expect market trends alone to be a good predictor of Rvola. Therefore, we find it surprising that cashtag variables persistently remain statistically significant, even in an environment where Covid-19 carries a lot of explanatory weight.

Our results give rise to questions about the informational content of cashtag messages, and the exact nature of the relationship between cashtags and markets. In the absence of sentiment data contained within these messages, we note that message volume could merely be a proxy for such data, and that sentiment could be a more precise predictor of volatility. Moreover, the information contained within cashtag messages may not be entirely unique, as it potentially mirrors information that is already conveyed through other media outlets. In which case, we would find that cashtag messages depict information arising from the media in general.

However, arriving at this conclusion would require further examination of the relationship, which is considered beyond the scope of this paper. Overall, our findings indicate that cashtag variables are associated with changes in volatility, and we find support in favor of H1.

6.2 Hypothesis 2

H2: Changes in company attention on Twitter, measured by message volume related to the company in general accumulated prior to trading hours, is associated with changes in volatility.

To test H2, we propose that company name variables are a suitable approximation of all messages related to the company in general.

6.2.1 Results

Table 6-2 contains regression results with coefficients for name variables.

We find that all Fe models produce significant coefficients for name variables on a 1 % level. However, the joint F-statistic reveal that none of the models are significant when the name variables are paired with entity fixed effects.

Results for HAR + Fe models show that name variable coefficients remain significant, except for L1.Nm and L1.Nm-sumred. The best HAR + Fe model contains Nm-17,5h. The coefficient is significant on a 10 % level, and is associated with an approximate increase in Rvola of 0,007 %-points when message volume increases by 100.

Interestingly, none of the HAR + Fe + Te models produce statistically significant name variables, and all coefficients turn negative. The change in polarity could indicate a spurious relationship between name variables and Rvola, as their coefficients are prone to change when additional control variables are introduced. In this case, time trends explain enough variation in Rvola to render name variables statistically insignificant, when paired with HAR and fixed effects.

We note that HAR + Fe + Nm-17,5h is the best model with respect to all name variables, as it produces the highest t-statistic in conjunction with a significant model test. The F-statistic is slightly lower than that of the models containing Nm-17,5h-sumred and Nm-24h-sumred, but not enough to affect overall assessment.

Table 6-2 Regression results for all company name variables

Base		[Twitter]	Coefficient	t-stat	F-stat
Fe	+	L1.Nm	0,0155***	3,30	0,80
HAR + Fe	+	L1.Nm	0,0027	1,16	112,68
HAR + Fe + Te	+	L1.Nm	-0,0006	-0,56	172,50
Fe	+	L1.Nm-sumred	0,0146***	3,45	0,84
HAR + Fe	+	L1.Nm-sumred	0,0023	0,95	113,13
HAR + Fe + Te	+	L1.Nm-sumred	-0,0013	-1,07	171,96
Fe	+	Nm-24h	0,0161***	3,20	0,79
HAR + Fe	+	Nm-24h	0,0046*	1,90	110,41
HAR + Fe + Te	+	Nm-24h	-0,0004	-0,32	172,79
Fe	+	Nm-24h-sumred	0,0125***	2,97	0,70
HAR + Fe	+	Nm-24h-sumred	0,0036*	1,72	111,12
HAR + Fe + Te	+	Nm-24h-sumred	-0,0013	-1,28	172,14
Fe	+	Nm-17,5h	0,0204***	2,94	0,72
HAR + Fe	+	Nm-17,5h	0,0067*	1,96	110,63
HAR + Fe + Te	+	Nm-17,5h	-0,0001	-0,07	173,99
Fe	+	Nm-17,5h-sumred	0,0158***	2,79	0,65
HAR + Fe	+	Nm-17,5h-sumred	0,0048*	1,72	111,01
HAR + Fe + Te	+	Nm-17,5h-sumred	-0,0016	-1,11	173,35

Fe refers to entity fixed effects, Te refers to time fixed effects. F-stat refers to the entire model.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

To test the robustness of HAR + Fe + Nm-17,5h, we run bootstrap estimations of the model. Sampling from 22 entity clusters yields an observed coefficient for Nm-17,5h equal to the initial estimation, significant on a 5 % level, with a 95 % confidence interval equal to $[0{,}0016; 0{,}0118]$. When sampling from 61 time clusters, we again observe the same coefficient for Nm-17,5h, significant on a 10 % level, with a 95 % confidence interval equal to $[-0{,}0003; 0{,}0137]$.

6.2.2 Discussion

Results from Table 6-2 reveal a trend where coefficients for name variables become less statistically significant as more control variables are introduced. Moreover, controlling for market volatility, through time fixed effects, renders all name variables insignificant. In accordance with our conclusion about cashtag variables, name variables that are closer to Rvola in time seem to have better predictive power. As opposed to cashtag variables, name variables are potentially subject to larger measurement error, as they might contain information that does not pertain to markets. In order to increase validity of such data, a subjective examination of popular terms for these companies would be necessary. We consider such an analysis to be beyond the scope of this paper and recognize that our systematic approach may inherently be flawed. However, we note that our own imposed assessments about search words could also introduce significant measurement errors, which could legitimize a systematic approach. Overall, the suggested name variables perform worse compared to cashtag variables. This could signal that our attention indicators are poorly specified, or simply that indicators relating to company name search words contain inadequate amounts of information relating to markets. Ultimately, we do not find name variables to carry unique explanatory power pertaining to volatility. Thus, we find insufficient evidence in support of H2. We note that controlling for time fixed effects is infeasible when attempting to predict future observations of volatility. Since name variables precede time fixed effects, they could still hold predictive power for volatility. Therefore, we continue to test H3 with our best performing name variable from H2.

6.3 Hypothesis 3

H3: Company attention can be utilized to improve the forecasting ability of volatility models.

To test H3, we propose our two best performing attention indicators derived from H1 and H2, Ct-17,5h and Nm-17,5h, in different combinations with the baseline model. Expanding upon previous correlation models, H3 will further examine the relationship between Twitter and volatility. To ease readability and manage space efficiently, we create short names for the models used in the forecasts: HAR + Fe (HAR), HAR + Fe + Ct-17,5h (HARct), HAR + Fe + Nm-17,5h (HARnm) and HAR + Fe + Ct-17,5h + Nm-17,5h (HARctnm).

6.3.1 In-sample

To conduct in-sample forecasting we estimate the models using the full sample, as portrayed in Table 6-3. Although differences vary, all three models improve upon the baseline model (HAR). Overall, the Ct-17,5h variable outperform Nm-17,5h, with HARct having a greater improvement in all performance measures compared to the baseline model. In terms of HARnm, the improvement from including Nm-17,5h appear negligible for MAPE and RMSE. However, the improvement of adding both Twitter variables seem to be greater than the sum of their parts, at least in terms of QLIKE and MAPE.

Table 6-3 Full sample: In-sample one-day ahead forecasting performance

Model	QLIKE	dQLIKE	MAPE	dMAPE	RMSE	dRMSE
HAR	0,1934		23,42 %		0,992 %	
HARct	0,1824	1,10 %	22,84 %	0,57 %	0,985 %	0,008 %
HARnm	0,1920	0,14 %	23,41 %	0,01 %	0,990 %	0,002 %
HARctnm	0,1803	1,31 %	22,78 %	0,64 %	0,983 %	0,009 %

Note: dQLIKE, dMAPE and dRMSE represent mean improvement from the baseline HAR-RV model in QLIKE, MAPE and RMSE, respectively.

In Table 6-4, we present the in-sample forecasting performance for period 1. Overall, results closely resemble the full in-sample forecast, with Ct-17,5h being the main contributing variable. We also note that the improvements for both HARct and HARctnm are greater than their full sample equivalent. Interestingly, dMAPE for HARnm is negative, which suggests that Nm-17,5h in some cases might reduce forecasting performance. Yet, the combination of both Twitter variables still yields the greatest improvement in all performance measures.

Table 6-4 Period 1: In-sample one-day ahead forecasting performance

Model	QLIKE	dQLIKE	MAPE	dMAPE	RMSE	dRMSE
HAR	0,1991		23,42 %		0,359 %	
HARct	0,1803	1,88 %	22,45 %	0,96 %	0,341 %	0,018 %
HARnm	0,1973	0,19 %	23,42 %	-0,01 %	0,359 %	0,001 %
HARctnm	0,1770	2,21 %	22,35 %	1,07 %	0,340 %	0,019 %

Note: dQLIKE, dMAPE and dRMSE represent mean improvement from the baseline HAR-RV model in QLIKE, MAPE and RMSE, respectively.

Table 6-5 presents the in-sample forecasting performance for period 2. Overall, performance measures indicate that Twitter variables carry less explanatory weight in period 2, compared to period 1, as improvements on the baseline model are more modest. To illustrate, we see that the largest improvement in QLIKE for period 2 is 0,28 %-points, compared to 2,20 %-points from period 1. HARctnm is consistently the best performer in the in-sample forecast. Additionally, we note that the level of QLIKE and MAPE for HAR is very similar in both periods, which indicates that the model is well specified with a break.

Table 6-5 Period 2: In-sample one-day ahead forecasting performance

Model	QLIKE	dQLIKE	MAPE	dMAPE	RMSE	dRMSE
HAR	0,1871		23,42 %		1,388 %	
HARct	0,1846	0,25 %	23,28 %	0,14 %	1,382 %	0,006 %
HARnm	0,1863	0,08 %	23,39 %	0,03 %	1,386 %	0,002 %
HARctnm	0,1840	0,31 %	23,25 %	0,17 %	1,380 %	0,008 %

Note: dQLIKE, dMAPE and dRMSE represent mean improvement from the baseline HAR-RV model in QLIKE, MAPE and RMSE, respectively.

6.3.2 Pseudo out-of-sample

In Table 6-6, we present the one-day ahead out-of-sample forecasting results. We combine forecasts conducted for all 13 subsamples into a single panel time series to test the full sample.

Table 6-6 Full sample: Pseudo out-of-sample one-day ahead forecasting performance

Model	QLIKE	dQLIKE	DM	MAPE	dMAPE	DM	RMSE	dRMSE	DM
HAR	0,2460			26,26 %			1,125 %		
HARct	0,2348	1,11 %	2,09**	25,69 %	0,57 %	2,32**	1,118 %	0,007 %	2,42**
HARnm	0,2432	0,27 %	1,88*	26,21 %	0,05 %	0,74	1,123 %	0,002 %	0,82
HARctnm	0,2314	1,46 %	2,70***	25,59 %	0,67 %	2,67***	1,117 %	0,008 %	2,35**

Note: dQLIKE, dMAPE and dRMSE represent mean improvement from the baseline HAR-RV model in QLIKE, MAPE and RMSE, respectively. DM represents the Diebold-Mariano test statistic for each dQLIKE, dMAPE and dRMSE, respectively. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The results in Table 6-6 reveal a similar trend to that which is described for the in-sample performance. Ct-17,5h is the main contributor to the forecasts, as illustrated by results for HARct. All improvements in performance measures for this model are statistically significant on a 5 % level. Similar to the full in-sample forecast, HARnm reveal modest improvements on the baseline model for all performance measures. Overall, HARctnm exhibits the lowest mean loss in all performance measures, and differences are statistically significant on a 1 % level for QLIKE and MAPE.

In Table 6-7, we present the forecasting performance for period 1. Nearly all models significantly improve upon the baseline, however, there is no significant improvement in MAPE for HARnm. Again, Ct-17,5h appear to be the main contributing variable, as evidenced by HARct, while Nm-17,5h performs better in conjunction with Ct-17,5h. In terms of mean improvements, HARctnm is the best performing model for period 1. The improvements are all statistically significant on a 5 % level. Finally, we note that HARct and HARctnm exhibit a greater improvement in period 1, compared to the full out-of-sample forecast.

Table 6-7 Period 1: Pseudo out-of-sample one-day ahead forecasting performance

Model	QLIKE	dQLIKE	DM	MAPE	dMAPE	DM	RMSE	dRMSE	DM
HAR	0,2170			24,39 %			0,394 %		
HARct	0,2002	1,68 %	1,70*	23,59 %	0,80 %	1,78*	0,381 %	0,013 %	1,95*
HARnm	0,2150	0,20 %	1,72*	24,41 %	-0,02 %	-0,28	0,392 %	0,001 %	1,72*
HARctnm	0,1965	2,05 %	2,08**	23,49 %	0,90 %	1,98**	0,379 %	0,014 %	2,23**

Note: dQLIKE, dMAPE and dRMSE represent mean improvement from the baseline HAR-RV model in QLIKE, MAPE and RMSE, respectively. DM represents the Diebold-Mariano test statistic for each dQLIKE, dMAPE and dRMSE, respectively. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 6-8 presents the results from the out-of-sample forecast performance for period 2. As expected, the contribution of Twitter variables is smaller compared to period 1. For period 2, we see that both QLIKE and MAPE are generally higher than for period 1. We see that HARctnm is the best performing model for period 2, and improvements in forecasting ability for HARct and HARctnm are both statistically significant. Interestingly, the improvement for HARnm is not statistically significant for any performance measure.

Table 6-8 Period 2: Pseudo out-of-sample one-day ahead forecasting performance

Model	QLIKE	dQLIKE	DM	MAPE	dMAPE	DM	RMSE	dRMSE	DM
HAR	0,2779			28,32 %			1,578 %		
HARct	0,2731	0,49 %	2,43**	28,00 %	0,32 %	2,45**	1,571 %	0,007 %	1,76*
HARnm	0,2744	0,35 %	1,18	28,19 %	0,13 %	1,08	1,575 %	0,003 %	0,69
HARctnm	0,2698	0,81 %	2,33**	27,90 %	0,42 %	2,39**	1,569 %	0,009 %	1,69*

Note: dQLIKE, dMAPE and dRMSE represent mean improvement from the baseline HAR-RV model in QLIKE, MAPE and RMSE, respectively. DM represents the Diebold-Mariano test statistic for each dQLIKE, dMAPE and dRMSE, respectively. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

6.3.3 Discussion

The results in H3 yield some interesting findings. Overall, the results are consistent across in-sample and out-of-sample forecasts. Analogous to results from H1 and H2, we find more evidence supporting the predictive power of the cashtag variable than the name variable. However, the combination of both Twitter variables consistently produces smaller mean losses. Crucially, we find the improvements from both models containing cashtags to be statistically significant in all out-of-sample tests. This is not the case for HARnm. As mentioned earlier, the company name variables are likely to contain more noise, and the erratic performance of HARnm across the out-of-sample tests seem to suggest the same. Yet, results from HARctnm indicate that Nm-17,5h does have merit.

Ideally, we would estimate the models on a larger sample, as outliers from period 2 would be given less weight. Instead, we include a break to ensure more consistent forecasts throughout each period. Breaks could compromise the generalizability of forecast results, as it requires knowledge about the distribution from which the forecasted value is drawn. However, there is no statistically significant change in coefficients for Twitter variables after the inclusion of a break for period 2, see Table 9-10 in Appendix 5. Hence, the introduction of a break should prove unproblematic for the forecasting ability of Twitter, and any inference drawn from this.

By testing forecasting performance for the chosen two periods, we can evaluate the models in both a stable environment and during a period of turmoil. Comparing MAPE and QLIKE losses in out-of-sample forecasts, we find that the overall forecasting is more precise in the stable period. When forecasting in a period of greater instability, one would expect lagged values of volatility to carry less explanatory weight. Hence, it is reasonable to think that the inclusion of exogenous variables, like Twitter variables, would yield a greater improvement in forecasting performance for period 2. This is consistent with findings for HARnm, although differences are not significant for period 2. Interestingly, the improvement in models containing cashtags are higher in period 1, which indicates that cashtags yield a greater improvement when forecasting in a stable environment. One possible explanation could be cross-panel correlation. As is well established, asset returns tend to become more correlated during periods of instability. This is also the case for our sample. If cashtags contain exclusive information relating to future realizations of volatility for a specific stock, it could be the case that this information is simply of less importance when stocks are so heavily affected by macroeconomic events.

Our findings reveal that the best performing model contain both Twitter variables, which is consistent for all out-of-sample forecasts. Interestingly, we find the correlation between Ct-17,5h and Nm-17,5h to be low, $\rho = 0{,}08$, which could indicate that each variable carries some unique explanatory power. Fitted values and residuals from regressing Rvola on time and entity fixed effects reveal a stronger correlation between fitted values and Nm-17,5h, while Ct-17,5h correlates more strongly with the residuals, see Table 9-11 and Table 9-12 in Appendix 6. Hence, we infer that Nm-17,5h relates more to systematic risk, while Ct-17,5h carry unique explanatory power pertaining to idiosyncratic risk. While previous findings from H2 indicated that name variables carried less explanatory power for realized volatility, these results shed new light on the nature of this relationship. Overall, we find support in favor of H3, in that company attention indicators from Twitter are useful for improving volatility forecasts. Further, we argue that both of our suggested indicators should be applied, as their informational content differ.

7 Concluding remarks

This paper provides evidence that changes in Twitter message volume is related to next-day changes in realized volatility, for 22 companies from S&P 100, and that the relationship can be expressed with attention indicators that rely on cashtags and search words for company names. In combination with an augmented HAR-RV model, we found variables containing aggregated message volume outside trading hours to be the best predictor of next day volatility. This suggests that Twitter information diffuses rapidly in markets, in line with the efficient market hypothesis. Further, using the best performing Twitter variables, we tested predictive ability by conducting forecasts and found that volatility is best predicted with a model containing both attention indicators. Overall, we found the predictive power of cashtag variables superior to that of name variables.

Surprisingly, we established that each variable explained different variation pertaining to volatility. Cashtags were found to explain variation relating to the idiosyncratic component of volatility, while company names explained variation in the systematic component. To our knowledge, our study is the first of its kind to uncover these key features about Twitter variables. Further, our results bring about questions regarding underlying mechanisms that dictate the relationship between Twitter and volatility. Such questions should attempt to answer the possibility of a latent social phenomenon. While our results are promising, we expect future research to include tools like sentiment analysis, to fully grasp the informational value contained within Twitter messages, as well as thorough examinations of search words to yield better results for name variables.

Due to Covid-19 and worldwide lockdowns, it is also possible that the usefulness of Twitter variables is exaggerated because of a general increase in user activity throughout our sample period. This would lead to artificially good predictors. To bypass these limitations, future research warrants a larger sample size, so that outlier events are given less weight in prediction models. Lastly, we suggest that more sophisticated spam-filtering is applied, as well as employing Twitter's Enterprise API for higher fidelity data, in order to reduce measurement error.

8 Bibliography

Aït-Sahalia, Y., Mykland, P. A., & Zhang, L. (2011). Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, *160*(1), 160–175. https://doi.org/10.1016/j.jeconom.2010.03.028

Andersen, T. G., & Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review*, *39*(4), 885–905. https://doi.org/10.2307/2527343

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, *96*(453), 42–55. https://doi.org/10.1198/016214501750332965

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

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, *59*(3), 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

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

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation Author (s): T . S . Breusch and A . R . Pagan. *Econometrica*, *47*(5), 1287–1294.

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, *28*(3), 591–605. https://doi.org/10.2307/1910133

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001

Cumby, R. E., & Huizinga, J. (1992). Testing the Autocorrelation Structure of Disturbances in Ordinary Least Squares and Instrumental Variables Regressions. *Econometrica*, *60*(1), 185–195. https://doi.org/10.2307/2951684

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

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

Engle, R. F., & Patton, A. J. (2007). What good is a volatility model? In *Forecasting Volatility in the Financial Markets* (pp. 47–63). Elsevier. https://doi.org/10.1016/B978-075066942-9.50004-2

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, *79*(2), 453–497. https://doi.org/10.3982/ECTA5771

Hentschel, M., & Alonso, O. (2014). Follow the money: A study of cashtags on Twitter. *First Monday*, *19*(8). https://doi.org/10.5210/fm.v19i8.5385

Levin, A., Lin, C. F., & Chu, C. S. J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, *108*(1), 1–24. https://doi.org/10.1016/S0304-4076(01)00098-7

Ma, F., Wei, Y., Huang, D., & Chen, Y. (2014). Which is the better forecasting model? A comparison between HAR-RV and multifractality volatility. *Physica A: Statistical Mechanics and Its Applications*, *405*, 171–180. https://doi.org/10.1016/j.physa.2014.03.007

McAleer, M., & Medeiros, M. C. (2008). Realized volatility: A review. *Econometric Reviews*, *27*(1–3), 10–45. https://doi.org/10.1080/07474930701853509

Müller, U. A., Dacorogna, M. M., Davé, R. D., Pictet, O. V, Olsen, R. B., & Ward, J. R. (1993). Fractals and intrinsic time: A challenge to econometricians. *Unpublished Manuscript, Olsen & Associates, Zürich*, 1–23.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent. *Econometrica*, *55*(3), 703–708.

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

Oliveira, N., Cortez, P., & Areal, N. (2013). Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13*, 1–8. https://doi.org/10.1145/2479787.2479811

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Patton, A. J., & Sheppard, K. (2015). Good Volatility, Bad Volatility: Signed Jumps and The

Persistence of Volatility. *Review of Economics and Statistics*, *97*(3), 683–697. https://doi.org/10.1162/REST_a_00503

Pesaran, M. H. (2004). General Diagnostic Tests for Cross Section Dependence in Panels. In *CESifo Working Paper Series No. 1229; IZA Discussion Paper No. 1240*. https://ssrn.com/abstract=572504

Pesaran, M. H. (2015). Testing Weak Cross-Sectional Dependence in Large Panels. *Econometric Reviews*, *34*(6–10), 1089–1117. https://doi.org/10.1080/07474938.2014.956623

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

Tafti, A., Zotti, R., & Jank, W. (2016). Real-time diffusion of information on twitter and the financial markets. *PLoS ONE*, *11*(8), e0159226. https://doi.org/10.1371/journal.pone.0159226

Thelwall, M. (2015). Evaluating the comprehensiveness of Twitter Search API results: A four step method. *Cybermetrics*, *18–19*(1), 1–10.

Twitter Inc. (2019a). *Q1 2019 Earnings Report*. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf?fbclid=IwAR1lfSxDvjxJUdBI5CgOmfJOEfNY2s--iUkH9nkF80WCVoZHsq_A24jbhB8

Twitter Inc. (2019b). *Selected Company Metrics and Financials*. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q4/Q4-2019-Selected-Financials-and-Metrics.pdf

Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, *100*(472), 1394–1411. https://doi.org/10.1198/016214505000000169

9 Appendices

This section contains complementary tables and figures for various chapters of this paper.

9.1 Appendix 1

Table 9-1 Two-Scales Estimator for an arbitrary asset on day t

TIME	PRICE	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	(all)
:	:	:	:	:	:	:	:
09:40	$u11$ $= \log(p11)$	$r_{t,i}^{k=1}$ $= u11 - u6$					$r_{t,i}^{(all)}$ $= u11 - u10$
09:41	$u12$		$r_{t,i}^{k=2}$ $= u12 - u7$				$r_{t,i+1}^{(all)}$ $= u12 - u11$
09:42	$u13$			$r_{t,i}^{k=3}$ $= u13 - u8$			$r_{t,i+2}^{(all)}$ $= u13 - u12$
09:43	$u14$				$r_{t,i}^{k=4}$ $= u14 - u9$		$r_{t,i+2}^{(all)}$ $= u14 - u13$
09:44	$u15$					$r_{t,i}^{k=5}$ $= u15 - u10$	$r_{t,i+3}^{(all)}$ $= u15 - u14$
09:45	$u16$	$r_{t,i+1}^{k=1}$ $= u16 - u11$					$r_{t,i+4}^{(all)}$ $= u16 - u15$
09:46	$u17$		$r_{t,i+2}^{k=2}$ $= u17 - u12$				$r_{t,i+5}^{(all)}$ $= u17 - u16$
09:47	$u18$			$r_{t,i+2}^{k=3}$ $= u18 - u13$			$r_{t,i+6}^{(all)}$ $= u18 - u17$
09:48	$u19$				$r_{t,i+2}^{k=4}$ $= u19 - u14$		$r_{t,i+7}^{(all)}$ $= u19 - u18$
09:49	$u20$					$r_{t,i+2}^{k=5}$ $= u20 - u15$	$r_{t,i+8}^{(all)}$ $= u20 - u19$
09:50	$u21$	$r_{t,i+2}^{k=1}$ $= u21 - u16$					$r_{t,i+9}^{(all)}$ $= u21 - u20$
:	:	:	:	:	:	:	:
		$RVar_t^{k=1}$ $= \sum_i (r_{t,i}^{k=1})^2$	$RVar_t^{k=2}$	$RVar_t^{k=3}$	$RVar_t^{k=4}$	$RVar_t^{k=5}$	$RVar_t^{(all)}$ $= \sum_i (r_{t,i}^{(all)})^2$

9.2 Appendix 2

Table 9-2 Unit-root test for rvola

Levin-Lin-Chu unit-root test for rvola

H0: Panels contain unit roots	Number of panels = 22
H1: Panels are stationary	Number of periods = 61
AR parameter: Common	Asymptotics: N/T -> 0
Panel means: Included	
Time trend: Not included	Cross-sectional means removed

ADF regressions: 1,59 lags average (chosen by AIC)

LR variance: Bartlett kernel, 12,00 lags average (chosen by LLC)

	Statistic	p-value
Unadjusted t	-16,7368	
Adjusted t*	-9,8333	0,0000

9.3 Appendix 3

Table 9-3 Correlation matrix

	Rvola	Rvolawk	Rvolamt	Dwkmt
Rvola	1,00			
Rvolawk	0,85	1,00		
Rvolamt	0,66	0,87	1,00	
Dwkmt	0,75	0,77	0,35	1,00

9.4 Appendix 4

Table 9-4 Serial correlation test

Cumby-Huizinga test for autocorrelation on residuals from the baseline HAR-RV model

H0: disturbance is MA process up to order q

H1: serial correlation present at specified lags >q

H0: q=0 (serially uncorrelated) | H0: q=specified lag-1

H1: s.c. present at range specified | H1: s.c. present at lag specified

lags	chi2	df	p-val	lag	chi2	df	p-val
1 - 1	0.223	1	0.6364	1	0.223	1	0.6364
1 - 2	7.315	2	0.0258	2	11.289	1	0.0008
1 - 3	10.454	3	0.0151	3	4.659	1	0.0309
1 - 4	14.115	4	0.0069	4	3.643	1	0.0563
1 - 5	16.546	5	0.0054	5	3.194	1	0.0739

Test robust to heteroskedasticity

Figure 9-1 Graphical representation of residuals

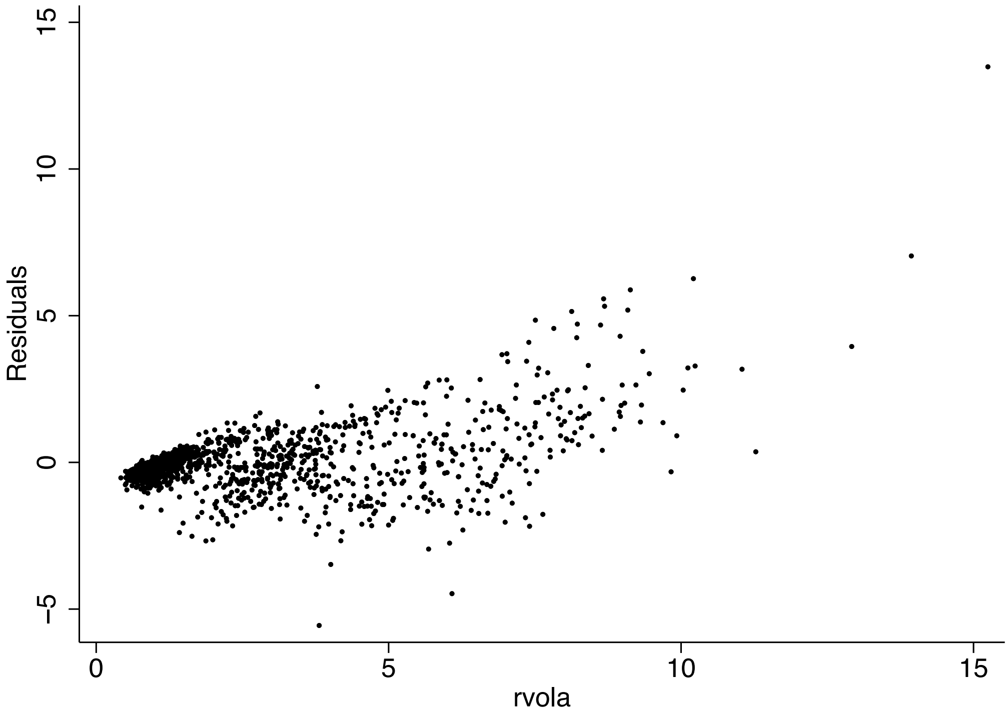


Table 9-5 Test for heteroskedastic residuals

Breusch-Pagan test for heteroskedasticity on residuals from baseline HAR-RV

H0: Constant variance

Variables: Fitted values of rvola

Chi2(1)	Prob > chi2
434,36	0,000

Table 9-6 Cross-panel correlation without controlling for time fixed effects

Test for cross-section independence on residuals without time fixed effects				
CD-test	p-value	Average joint T	mean ρ	mean abs(ρ)
67,33	0	61	0,57	0,57

Notes: Under the null hypothesis of cross-section independence, CD ~ N(0,1)

P-values close to zero indicate data are correlated across panel groups.

Table 9-7 Cross-panel correlation with time fixed effects

Test for cross-section independence on residuals with time fixed effects				
CD-test	p-value	Average joint T	mean ρ	mean abs(ρ)
-4,213	0	61	-0,04	0,19

Notes: Under the null hypothesis of cross-section independence, CD ~ N(0,1)

P-values close to zero indicate data are correlated across panel groups.

Table 9-8 OLS regression with bootstrapped SEs from replications based on 22 clusters in entity

rvola	Observed coef.	Bootstrap Std. Err.	z	P > | z |	Normal-based 95% conf. Interval	
L1.Rvolac	0,427***	0,064	6,64	0,000	0,301	0,553
Rvolamtc	0,373***	0,056	6,66	0,000	0,264	0,483
Dwkmt	0,638***	0,082	7,80	0,000	0,478	0,799
Constant	2,421***	0,152	15,95	0,000	2,123	2,718
α_i	-		(83,74***)	(0,000)		
Observations	1342					
Wald chi2(24)	3166,61					
Prob > chi2	0,000					

Joint Chi2-statistic for α_i and corresponding p-value in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9-9 OLS regression with bootstrapped SEs from replications based on 61 clusters in time

rvola	Observed coef.	Bootstrap Std. Err.	z	P > | z |	Normal-based 95% conf. Interval	
L1.Rvolac	0,427***	0,107	3,98	0,000	0,217	0,637
Rvolamtc	0,373***	0,116	3,23	0,001	0,306	0,600
Dwkmt	0,638***	0,173	3,68	0,000	0,299	0,979
Constant	2,421***	0,160	15,14	0,000	2,107	2,734
α_i	-		(419,07***)	(0,000)		
Observations	1342					
Wald chi2(24)	968,61					
Prob > chi2	0,000					

Joint Chi2-statistic for α_i and corresponding p-value in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

9.5 Appendix 5

Table 9-10 Regression results: HARctnm with break

rvola	Coef.	Newey-West Std. Err.	t	P > | t |	95% conf. Interval	
L1.Rvolac	0,189***	0,065	2,92	0,004	0,062	0,317
Rvolamtc	0,376*	0,211	1,78	0,076	-0,039	0,790
Dwkmt	0,036	0,098	0,37	0,715	-0,157	0,229
Ct-17,5h	0,173***	0,037	4,66	0,000	0,010	0,245
Nm-17,5h	-0,002	0,004	-0,60	0,547	-0,009	0,005
L1.Rvolac * Break	0,161**	0,075	2,16	0,031	0,015	0,308
Rvolamtc * Break	-0,162	0,217	-0,74	0,457	-0,589	0,265
Dwkmt * Break	0,590***	0,114	5,16	0,000	0,366	0,815
Ct-17,5h * Break	-0,371	0,063	-0,59	0,558	-0,161	0,087
Nm-17,5h * Break	0,007	0,006	1,20	0,231	-0,004	0,018
Break	1,061***	0,254	4,18	0,000	0,563	1,559
Constant	1,799***	0,338	5,320	0,000	1,136	2,463
α_i	-		(3,84)	(0,000)		
Observations	1340					
Wald chi2(24)	128,93					
Prob > chi2	0,000					

Joint F-statistic for α_i and corresponding p-value in brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

9.6 Appendix 6

Regression of Rvola on entity (α_i) and time (λ_t) fixed effects described in equation (18).

$$Rvola_{i,t} = \beta_0 + \alpha_i + \lambda_t + u_{i,t} \tag{18}$$

Fitted values ($\widehat{Rvola}_{i,t}$) and residuals ($\hat{u}_{i,t}$) are saved and used in regression with Ct-17,5h and Nm-17,5h.

Table 9-11 Regression of fitted values on Twitter variables

$\widehat{Rvola}_{i,t}$	Coefficients
Ct-17,5h	0,066
	(1,09)
Nm-17,5h	0,019***
	(3,91)
Constant	2,588***
	(34,23)
Observations	1340
F-test	8,67
Prob > F	0,002

t-statistics in parentheses and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9-12 Regression of residuals on Twitter variables

$\hat{u}_{i,t}$	Coefficients
Ct-17,5h	0,105***
	(4,78)
Nm-17,5h	-0,001
	(-0,71)
Constant	-0,071***
	(-2,58)
Observations	1340
F-test	8,67
Prob > F	0,002

t-statistics in parentheses and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$