The effect of non-financial Twitter sentiment on stock price volatility

A thesis submitted in the partial fulfillment of the requirements for the degree of Master in Quantitative Finance and Actuarial Sciences

Tilburg School of Economics and Management

by

Jonas de Boo (u1272278)

Supervisor:

Dr. P. Cizek

Final Version - November 30, 2021

Abstract

This paper explores the in-sample effects of Twitter on the stock price return volatility of 12 major U.S. companies. Motivated by the increased attention among retail investors, the importance of social media in our daily lives, and the adoption of non-financial practices in investment strategies, this study only considers non-financial tweets. To determine the Twitter-based sentiment score, this research employs a domain adjusted variant of the VADER lexicon. This paper develops a new method to aggregate tweets based on readily available *attention variables*. Besides the sentiment, this study considers both the number of tweets and the number of interactions to account for the effect of Twitter on stock price return volatility. To describe the return process, this research uses an auto-regressive moving-average structure with apARCH innovations. Both the asymmetric effects and the effects of the volatility of the Twitter variables on the stock price return volatility are investigated. It is found that in a constant parameter setting, the volatility of the Twitter variables can significantly explain the volatility of the stock price returns. In general, Twitter only affects one day ahead volatility. To account for possible varying effects, a smooth, macroeconomic dependent function is employed to investigate whether the effect of Twitter is subject to the state of the economy. The effects of Twitter are concluded to be dependent on the overall state of the economy. During economic downturn, Twitter successfully explains the conditional volatility of the majority of the companies. Moreover, it is concluded that negative sentiment has a larger effect on volatility than positive sentiment.

Contents

1 Introduction

> "The beauty of social media is that it will point out your company's flaws; the key questions is how quickly you address these flaws."
>
> *Erik Qualman (Socialnomics: How Social Media Transforms the Way We Live and Do Business (ed. 2009))*

Over the years, numerous *micro-blogging* platforms have gained a prominent role in the daily conversations of individuals. In particular, over the last decade, Twitter has grown out to be more and more of a news and actualities sharing platform (Kwak, Lee, Park, & Moon, 2010). People post *tweets*, small messages containing up to 280 characters, which can be read by their followers, sometimes attaching *hashtags* referring to particular threads or topics.

On Twitter, users can continuously display their feelings and emotions, possibly influencing others into actions. This makes Twitter a great source of *sentiment mining*, which is the practice of deriving sentiment scores based on the tweets posted on the platform. The rise of the role of Twitter in our daily lives has happened gradually along with the rise of computational power. This has increased the popularity of Natural Language Processing (NLP) techniques that analyse textual data to find relevant structures (*e.g.* the spam filter in our e-mail inbox). Moreover, an increased amount of financial news is shared on Twitter, which extends the number of financial applications. For instance, Dredze, Kambadur, Kazantsev, Mann, and Osborne (2016) find that Twitter can be regarded as a valuable source of financial information. This allows for new behavioural finance research, which investigates the relationship between psychology and financial markets, and in most cases relaxes the assumptions that investors behave rationally at all time. Although using Twitter to model financial market performance has been done in numerous research (Audrino, Sigrist, & Ballinari, 2020; Fan, Talavera, & Tran, 2020; Ramco, Aleksovski, Calderelli, Grcar, & Mozetic, 2015), little attention has been paid to the type of tweets included in these researches. Baurichter (2021), finds that retail investors account for the majority of the stock price movements, and additionally, that these investors are guided more by their emotion than institutional investors. In their article, Rakowski, Shirley, and Stark (2021) discuss the effect of Twitter and observe a significant effect, especially for stocks primarily traded by retail investors. Following the arguments of Da, Engelberg, and Gao (2011) and Hsu, Lu, and Yang (2021), it is argued that retail investors react differently to news than institutional investors, and are therefore more likely to be affected by social media coverage. Additionally, it is argued that after the recent GameStop mania, retail investors are likely to influence stock market prices for longer periods to come (Fitzgerald, 2021). Therefore, this paper considers tweets that arguably affect retail investors to a greater extent.

Over the past years, the interest in responsible investing has surged. Practices like Corporate Social Responsibility (CSR) or Environmental, Social and Governance (ESG) have been increasingly integrated into the financial world. The rising popularity of these terms associated with the way corporations should behave according to the public, gives rise to the question whether investors are likely to be influenced by announcements regarding these terms. As the quote illustrates, it can be argued that social media can serve as an accelerator for thoughts or opinions towards companies. From the news article 'Deliveroo crashes in IPO after remarkable run-up'[1], it can be seen that the interest of institutional investors in the shares of Deliveroo descreased because of the bad working conditions at Deliveroo. This supports the intuition that investors are more concerned with responsible investments. From Da et al. (2011), it is clear that retail investors react different to news announcements than more sophisticated institutional investors. Combining these findings, it is proposed that non-financial Twitter coverage can be used to predict stock price return

[1] NOS: Deliveroo onderuit bij beursgang na opmerkelijke aanloop

volatility. The volatility of the stock price returns is described as the degree to which the returns vary over time. Therefore, this paper explores whether Twitter drives stock price fluctuations. It is investigated whether volatility models that take into account Twitter-variables, capture the volatility of the stock price returns better than models that do not account for these variables. Simultaneously, this paper is concerned with exploring the effect of Twitter coverage that is argued to predominantly affect retail investors. Since the U.S. stock market is the most prominent and the U.S. has the most active Twitter users[2], U.S. data is used.

Since Engle (1982) developed the Autoregressive Conditional Hetereskedascity (ARCH) and Bollerslev (1986) proposed the Generalized ARCH (GARCH) model, modeling the volatility of stock prices, inflation rates or GDP growth has been a hot topic in finance. Investors dislike the exposure to heavy swings of their stock portfolio, similarly countries do not like a very fluctuating inflation rate. In most cases, higher volatility is associated with increasing levels of risk. Since investors try to balance their risks and returns high volatility causes constant portfolio reallocation (Markowitz, 1952). Furthermore, Engle and Patton (2001) discuss the fact that throughout the financial world, precise volatility forecasts are crucial. Since inaccurate volatility models can lead to excessive risk taking, the model from Bollerslev (1986) has been under constant revision, with researchers constantly proposing adjustments to account for additional effects in financial time series. Black (1976) introduced the phenomenon of the *leverage effect*, which proposes that volatility tends to rise in periods when excess returns are lower than they were expected, and falls when excess returns exceed their expectation. Additionally, Schwert (1989) reports increasing levels of stock volatility in times of economic distress or war, which can only partly be explained by increasing market volatility. To this extent, asymmetric volatility models that take into account the leverage effect are proposed by Glosten, Jagannathan, and Runkle (1993) and Nelson (1991). In these models, negative shocks can have a larger effect than positive shocks.

More recent is the introduction of exogenous variables into stock price volatility models, which are of interest when assessing effects of non-financial Twitter coverage. Engle and Patton (2001) show the significance of the inclusion of the T-bill rate when making forecasts of the Dow Jones Industrial Index returns volatility. Lately, Hsu et al. (2021) proposed to include sentiment variables in the volatility equation and report significant effects of these exogenous variables. Additionally, in the recent article by de Winter and van Dijk (2021), the authors found a significant relationship between headlines in the Dutch Financial Times (Het Financieele Dagblad) and macroeconomic performance. By accounting for these headlines, short term volatility forecast error was reduced by up to 10%. These articles underline the observed significance of accounting for sentiment and news based effects in volatility modeling.

Furthermore, Bollerslev and Ghysels (1996) consider an extension of the GARCH model, which allows for the parameters to depend on cyclical patterns. In this article, it is found that these types of varying parameter models increase the flexibility of the GARCH models compared to the time-invariant GARCH models. Additionally, allowing for the existence of structural breaks or time-varying coefficients in volatility models have been proven to greatly improve volatility forecasts (Andreou & Ghysels, 2002). Amado and Teräsvirta (2008) introduce a time-varying GARCH model, where the transition between the multiple regimes is modeled by a smooth transition function. Considering the fact that the effect of Twitter may also vary through time, this specification can greatly improve existing methodologies that incorporate Twitter sentiment via a single regime model.

As illustrated by Engle and Patton (2001), accurate volatility models provide great benefits across the entire financial industry. Since social media is increasingly important in the way that we interact with

[2]Number of Twitter users per country

others, recommend (financial) products and display our (dis)satisfaction with everyday events, the role of Twitter can not be disregarded when investigating stock price movements. Especially in the case of individual investors, it can be argued that Twitter plays a significant role in their news dissemination. Therefore, this paper will investigate if non-financial Twitter data can be used to explain stock price volatility. In particular, this research aims to answer how Twitter exerts influence on volatility, when Twitter influences volatility, and how long Twitter affects volatility. To answer this question, multiple models that combine common features are constructed to determine which effect Twitter exerts on volatility.

This analysis is performed in-sample, and mainly focuses on discovering the observed differences in the impact of sentiment through time for different companies, as suggested by Audrino et al. (2020). However, uncovering in detail the ways that Twitter activity influences stock price volatility can be used to improve out-of-sample volatility forecasting in future research. Moreover, this analysis will be useful to determine which content shared on Twitter that mainly drives stock price volatility

Using a combination of the proposed methodologies, the question how and when Twitter coverage affects stock price volatility can be answered more precisely. To this extent, multiple models are investigated across a sample of U.S. multinationals and compared to determine how Twitter coverage can successfully explain volatility. Additionally, to determine what kind of Twitter information affects volatility, this research specifically selects tweets following a manually constructed *ESG*-dictionary, which contains multiple terms linked to non-financial performance. Especially since these interests are sometimes conflicting, creating more insights is very valuable. For instance, when Apple invests ten million in creating eco-friendly offices, this is likely to be good for the environment, however unlikely to increase sales.

Earlier research has paid very little attention to the selection of an appropriate sentiment classification method. These articles made use of sentiment classification methods that have been shown to perform not very accurate in determining the appropriate sentiment. Especially since more advanced sentiment classification methods - which are proven to be very accurate - are performing increasingly better due to the increasing availability of data, paying more attention might turn out to be valuable (Go, Bhayani, & Huang, 2009). Determining the correct sentiment from textual data is crucial in these types of research. Since deriving the incorrect sentiment score might result in failure to discover meaningful links between sentiment and volatility. Finally, a parsimonious method that can be used to derive a weighted average of multiple sentiment scores is developed. This improves earlier research where calculated sentiment scores are simply averaged (Fan et al., 2020; Ramco et al., 2015), and takes into account the relative importance of tweets in the conversations that are happening on Twitter.

The remainder of this thesis is structured as follows. Section 2 introduces the basic GARCH model along with common estimation techniques. Additionally, this section explains the rationale behind common GARCH alternatives and introduces the necessary assumptions associated with these alternatives. Then, it section 3.1 explains how to leverage Twitter to determine the impact of tweets with non-financial content, and section 3.2 introduce various sentiment classification methods. Section 3.3 introduces a model that calculates the weighted average sentiment score. In section 4, the proposed econometric models are explained, along with the appropriate assumptions, estimation procedure, and tests. Additionally, this section presents a time-varying specification, which is employed to determine in more detail the periods during which Twitter exerts significant influence on volatility. Section 5 discusses the application of the stated theory to empirical applications, provides the summary statistics, and discusses several time series plots. The results are analysed and shown in section 6. Lastly, section 7 summarizes the contents of this research and section 8 provides concluding remarks and recommendations for future research.

2 Preliminaries

This research considers various methodologies to determine the effect of non-financial Twitter-based public sentiment on stock price movements. To provide the reader with some intuition about how Twitter variables can enter volatility models, this chapter introduces common volatility modeling techniques. Moreover, this chapter provides the necessary assumptions and estimation techniques associated with these models.

2.1 The GARCH model

Since Engle (1982) proposed the Autoregressive Conditional Heteroskedascity (ARCH) model to identify stock price volatility, various extensions to this model have been proposed. In the first place, Bollerslev (1986) introduced the Generalized ARCH (GARCH) model, a natural generalization of the model proposed by Engle (1982). In his article, Bollerslev (1986) first provides the specification of the simplest GARCH(1,1) process, which is later extended to the general GARCH(p, q) to allow for longer memory and more flexible lag structure. To define the GARCH(1,1) process, let ε_t be a real-valued discrete time stochastic process with information available up to time t

$$\varepsilon_t = \sigma_t z_t \tag{1}$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{2}$$

where $\omega > 0$, $\alpha \geq 0$ and $\beta \geq 0$ to ensure non-negativity of the process σ_t. Note that the processes z_t and σ_t are unobservable and can only be estimated. The variables z_t are called the *innovations* and are assumed to be independent and identically distributed with zero mean and unit variance, *i.e.* $E[z_t \,|\, \varepsilon_{t-1}, \ldots, \varepsilon_1] = 0$ and $E[z_t^2 \mid \varepsilon_{t-1}, \ldots, \varepsilon_1] = 1$. Throughout this research, σ_t^2 denotes the conditional variance of ε_t given information up to time t, let $\mathcal{F}_{t-1}$ denote this information set including all information up to time t. In case of the model from equation (2), this set is equal to $\mathcal{F}_{t-1} = \{\varepsilon_u, u < t\}$. Then, the assumption on the innovations is rewritten to $E[z_t \mid \mathcal{F}_{t-1}] = 0$ and $E[z_t^2 \mid \mathcal{F}_{t-1}] = 1$.

To guarantee stationarity of the conditional variance process from equation (2), it must hold that $\alpha+\beta < 1$ (Bollerslev, 1986), which means that the first moment and autocorrelations are time-invariant. In many empirical research, GARCH processes are used to determine how long shocks to the process described in equation (2) affect future values of the conditional volatility (Nelson, 1990), which is called the *persistence*. As the conditional volatility rises after shocks of ε_{t-1}, it is of interest how long these shocks keep affecting the conditional volatility many periods ahead. In the GARCH(1,1)-case, the total degree to which shocks 'persist', is measured by the sum $\alpha + \beta$, with the sum being larger indicating a higher degree of persistence. As Lamoureux and Lastrapes (1990) point out, it is generally observed that when the GARCH model is applied to high frequency financial data, shocks to variance persist for longer periods ahead in time and the conditional volatility only slowly decays.

Estimation of the parameter values in equation (2) is commonly done by the Maximum Likelihood Estimator (MLE). Based on a likelihood function, these are parameter estimates that make the observations of the conditional volatility the most 'likely' based on the distribution of the innovations z_t. For computational convenience, in most cases the log likelihood function $\log \mathcal{L}(\varphi)$ is maximized (Bain & Engelhardt, 1987). In the initial article by Bollerslev (1986), the innovations are assumed to follow a standard normal distribution. However, this assumption does generally not hold, as financial time series commonly have more fat-tailed distributions (Feng & Shi, 2017). This common attribute is also observed by Bollersev and Wooldridge (1992), who disregard the restrictive normality assumption. The authors proceed by introducing the asymptotic efficiency of a Quasi Maximum Likelihood Estimator (QMLE) that estimates the parameters in a consistent and efficient way. The QMLE differs from the traditional Maximum Likelihood Estimator in that the QMLE allows some distributional assumptions to be misspecified (White, 1982). When the quasi maximum likelihood function is not oversimplified, the parameters that maximize this function are still consistent parameter estimates of the true parameter values (Bollersev & Wooldridge, 1992).

The generalization of the GARCH(1,1) model, the GARCH(p, q) specification tackles the empirical problems observed by using the ARCH model from Engle (1982), as a relatively long lag in the conditional variance equation is often needed. Let p and q denote the lag order of the process ε_t and σ_t, respectively, then

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2, \tag{3}$$

where $\omega > 0$, $\alpha_i \geq 0$ for $i = 1, \ldots, p$ and $\beta_i \geq 0$ for $i = 1, \ldots, q$ to ensure non-negativity. Again, the same assumptions on the innovations hold, that is $E[z_t|\mathcal{F}_{t-1}] = 0$ and $E[z_t^2|\mathcal{F}_{t-1}] = 1$, where $\mathcal{F}_{t-1} = \{\varepsilon_u, u < t\}$. To guarantee time invariant first moments and autocorrelations, it must hold that $\sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{q} \beta_i < 1$ (Bollerslev, 1986). However, by thorough empirical research, it is commonly argued that in financial applications, the model with $p = q = 1$ provides a generally good fit on the time-series data. Hansen and Lunde (2005) investigate the use of over 330 GARCH-type models, with varying levels of lags. In their conclusion, the authors argue that no model succeeds in beating the standard GARCH$(1, 1)$ model. Therefore, the GARCH(1,1) model serves as a benchmark model in the following alternatives.

2.2 Extensions of the standard GARCH$(1, 1)$ model

In this subsection, various common extensions of the GARCH$(1, 1)$ model are discussed and explained. These extensions on the previously introduced idea by Bollerslev (1986) provide alternatives that have been tested extensively through empirical research, and that allow for the inclusion of Twitter-based variables into the equation from equation (2). These methods provide the groundwork to develop a model that allows for inclusion of Twitter-based sentiment variables into the standard model from equation (2).

The ARMA(P, Q)-GARCH model

In his article, Weiss (1984) discusses a special case of the ARCH model by Engle (1982), which models a variable y_t as an Auto Regressive Moving Average (ARMA) process with ARCH errors. As mentioned by Franq and Zakoian (2004), it is common in financial applications to use ARMA-type models to fit the conditional mean of the return process, and let the errors of this process follow a GARCH process. This allows return time series to incorporate a drift term, autocorrelations, or seasonal effects in the conditional mean equation. Similar to the model from equation (3) the ARMA model has a lag structure for both parts. The first part of the ARMA(P, Q) model corresponds to the autoregressive (AR) part, where lagged values of y_t up to lag P enter the equation of y_t. The second part incorporates the moving-average (MA) structure, which describes y_t as a function of the Q lagged estimation errors $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-Q}$. This leads to the following specification of the conditional mean process of y_t:

$$\begin{aligned} y_t &= \mu + \sum_{i=1}^{P} \gamma_i y_{t-i} + \sum_{i=1}^{Q} \delta_i \varepsilon_{t-i} + \varepsilon_t, \\ &= \mu + \sum_{i=1}^{P} \gamma_i L^i y_t + \left(1 + \sum_{i=1}^{Q} \delta_i L^i\right) \varepsilon_t, \end{aligned} \tag{4}$$

where L denotes the lag operator, *i.e.* $L^i y_t = y_{t-i}$ and $L^i \varepsilon_t = \varepsilon_{t-i}$. Let $\mathcal{G}(L) = 1 - \sum_{i=1}^{P} \gamma_i L^i$ and $\mathcal{D}(L) = 1 + \sum_{i=1}^{Q} \delta_i L^i$ denote the AR and MA polynomials respectively, then equation (4) is equal to $\mathcal{G}(L)y_t = \mu + \mathcal{D}(L)\varepsilon_t$. To guarantee stationarity and invertibility of the ARMA(P,Q) process, it must hold that the roots of $|\mathcal{G}(L)| = 0$ and $|\mathcal{D}(L)| = 0$ are outside the unit circle (Ling & McAleer, 2003). To derive the ARMA-GARCH specification, the errors ε_t of equation (4) are assumed to follow the process described by equation (2).

Asymmetric GARCH models

In their article Ding, Granger, and Engle (1993) argue that there is no specific reason to model the conditional variance process as a linear function of lagged squared residuals. Considering the *leverage effect* reported by Black (1976), it is argued that not only the level, but also the sign of shocks to the values ε_t have an effect on the conditional volatility.

The most commonly used GARCH alternatives that allow for different effects of negative or positive shocks of $\varepsilon_1, \ldots, \varepsilon_T$ on the conditional volatility are the GJR-GARCH from Glosten et al. (1993) and the Exponential GARCH model of Nelson (1991). As the statistical behaviour of the EGARCH model estimates is not readily available under general conditions (Franq & Thieu, 2019), this research employs the GJR-GARCH model to take into account the asymmetric effects of positive versus negative residuals. The GJR-GARCH(1,1) model is defined as follows

$$\sigma_t^2 = \omega + \alpha_i^* \varepsilon_{t-1}^2 + a I_{t-1} \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{5}$$

where I_{t-1} is an indicator function, which takes value one when $\varepsilon_{t-1} < 0$ and zero otherwise. To ensure non-negativity of the conditional variance process, $\omega > 0$, $\alpha^* \geq 0$, $a \geq 0$ and $\beta \geq 0$. Again, let ε_t be characterized by equation (1), and let z_t be an i.i.d. sequence of innovations. Additionally, assume $E[z_t \mid \mathcal{F}_{t-1}] = 0$ and $E[z_t^2 \mid \mathcal{F}_{t-1}] = 1$, where $\mathcal{F}_{t-1} = \{\varepsilon_u, u < t\}$.

Since the GJR-GARCH model as described above is a special case of the asymmetric power ARCH (apARCH) model by Ding et al. (1993)[3], the notation of the apARCH with $\kappa = 2$ is used throughout this research. Define this model by

$$\sigma_t^2 = \omega + \alpha \left(|\varepsilon_{t-1}| - \psi \varepsilon_{t-1}\right)^2 + \beta \sigma_{t-1}^2, \tag{6}$$

where $-1 < \psi < 1$ captures the possible asymmetric effects of shocks of ε_t to the conditional volatility. Similar to earlier definitions, $\omega > 0$, $\alpha \geq 0$ and $\beta \geq 0$ to guarantee non-negativity of σ_t. When the distributions of the innovations is symmetric, Ling and McAleer (2002) suggest that in order for the process of equation (6) to be stationary, it must hold that $\alpha(1-\psi)^2 + \alpha(1+\psi)^2 + \beta < 1$, which is equivalent to $\alpha(1+\psi^2) + \beta < 1$. Similar to the GARCH(1,1) case, $\alpha(1+\psi^2) + \beta$ also denotes the degree to which shocks to the conditional volatility persist, and have a lasting effect on future conditional volatility values.

From equation (6) it can be seen that for positive values of ψ, negative values of ε_t have more effect than positive values on the conditional volatility, which is in line with the observations by Black (1976) and Schwert (1989). On the contrary, for negative values of ψ, positive values of ε_t have a larger effect on the conditional volatility.

The GARCH-X model

Multiple empirical research articles investigate the use of economic or financial indicators in GARCH-type models, but also the inclusion of exogenous sentiment variables (see Hsu et al., 2021 or Rupande,

[3] As shown by Ding et al. (1993), when $\kappa = 2$, the apARCH model can be rewritten to the GJR-GARCH model of equation (5). That is, when $0 \leq \psi_i < 1$ it can be shown that $\alpha_i^* = \alpha_i(1-\psi_i)^2$ and $b_i = 4\alpha_i\psi_i$ and when $-1 < \psi < 0$ the parameters of equation (5) are $\alpha_i^* = \alpha_i(1+\psi_i)^2$ and $b_i = -4\alpha_i\psi_i$.

Muguto, and Muzindutsi, 2019 for recent examples). These models are generally refered to as GARCH-X models and allow for the inclusion of exogenous variables into the conditional variance equation. Han and Kristensen (2014) provide a specification where lagged values of the exogenous variable x_t enter the conditional volatility equation of equation (3). As discussed, throughout this research it is assumed that $p = q = 1$, however, it is interesting to include up to the R-th order lag of the exogenous variable x_t, leading to the following GARCH-X specification

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \sum_{i=1}^{R} \pi_i x_{t-i}^2, \tag{7}$$

where $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ and $\pi \geq 0$ to ensure non-negativity of the conditional variance process. To guarantee stationarity of σ_t, it must hold that $\alpha + \beta < 1$, as shown by Bollerslev (1986). To succesfully include the exogenous variable x_t into the conditional variance equation, it must hold that the values $x_1, \ldots, x_T$ form a stationary time series process. Following Han and Kristensen (2014) the covariates are required to be exogenous in the sense that $E[z_t \,|\, \mathcal{F}_{x,t-1}] = 0$ and $E[z_t^2 \,|\, \mathcal{F}_{x,t-1}] = 1$, where the information set is defined by $\mathcal{F}_{x,t-1} = \{\varepsilon_u, x_u, u < t\}$. This assumption limits the possibility of x_t, as choosing a stock return will likely violate the exogeneity assumption.

Similar to the GARCH(1,1) models, the parameters of the proposed alternatives can be estimated by the QMLE. In their article, Ling and McAleer (2003) provide the asymptotic consistency of the QMLE of the ARMA-GARCH specification, which is then shown again under less restrictive assumptions by Franq and Zakoian (2004). Han and Kristensen (2014) show the consistency of the QMLE in the GARCH-X setting, and Franq and Thieu (2019) provide the conditions for consistency of the QMLE for the apARCH-X model. Section 4 discusses in more detail the conditions that must hold for consistency of the QMLE, introduces the quasi likelihood function and provides the models that are used in this paper.

Additionally, section 4 introduces a robustness check to test whether the models can be extended to allow for time-varying effects. This can help to determine during what periods Twitter can be leveraged to accurately model conditional volatility. Furthermore, Lamoureux and Lastrapes (1990) provide arguments that when a model is unable to signal shifts, it is likely to overstate the true level of persistence.

3 Modeling sentiment

Section 2 discussed what types of models can be used to model stock price movements, and how exogenous variables might be included in these models. This chapter focuses on how Twitter can be leveraged to determine non-financial public sentiment towards U.S. companies, to determine its effect on stock price volatility. This platform is argued to play a pivotal role in the information and opinion stream that Twitter users experience. This chapter discusses techniques that handle the classification of sentiment based on textual data (the selected *tweets*). Furthermore, this chapter discusses how to combine the sentiment score of multiple tweets within one time period to derive the total sentiment score. Lastly, other variables that might be valuable in assessing the total sentiment that can be derived based on a Twitter dataset are highlighted.

3.1 Twitter data

To retrieve the necessary tweets, Twitter provides researchers the option to find tweets via the Academic License of their Twitter search API. Via the Academic License of the Twitter API it is fairly easy to get historical tweets obeying the restrictions imposed by the user. Furthermore, any historical tweet dating back to as early as 2006 can be found, extending the possibilities of the regular license, which only allows retrieval of tweets dating back up to one week. As this license was only deployed in 2020, this extends the number of opportunities with regard to Twitter-based sentiment analysis.

Constructing the *search query*

Since randomly retrieving tweets does not provide an overview of the non-financial public sentiment towards the selected U.S. corporations, the Twitter API allows the user to specify a *search query.* This search query, consisting of at most 1024 characters, holds the specifications of the content (*i.e.* words, language or time-interval) that must hold for all tweets. In order to retrieve tweets that are valuable in the assessment of non-financial public sentiment, the following is done. First, a list of all Environmental, Social and Goverance (ESG) related words is created manually, containing words that are affiliated with non-financial performance of companies. This list can be found in appendix A.1.1 in table 12. This list has been constructed with help of the ESG glossary of Allianz[4], completed with multiple additions describing the non-financial actions of companies (*i.e.* fraud, pollution or slavery). By specifying these words, the Twitter API only retrieves tweets that contain at least one word from this *ESG dictionary.*

Besides containing one or more words from the *ESG dictionary*, a tweet must contain either the company name, official company name or the company stock exchange ticker symbol accompanied with a so-called cashtag ($), *e.g.* tweets must contain either Apple, Apple Inc. or $AAPL when retrieving tweets for Apple.

As the search query can also hold terms that filters out tweets with content that is not relevant, a *negation dictionary* is created, which limits the amounts of random tweets retrieved. This negation dictionary is company-specific and makes sure that words from this dictionary do not occur in the retrieved tweets. Creating this negation dictionary is done manually by inspecting a small subsample of the retrieved tweets, and looking for tweets that are evidently spam or are not of interest in this research, such as product offers. This negation dictionary is crucial as it filters out a lot of tweets that do not contribute to the goal of this research. Besides filtering the historical tweets based on the *ESG dictionary* and *negation dictionary*, several other constraints are imposed on the retrieval of the relevant tweets. Tweets in this research are only written in English, are not promoted (commercial) tweets and are not retweets, quotes, or replies to other tweets.

[4] Allianz ESG glossary

Public metrics

Besides the content of the retrieved tweets, the Twitter API also allows to retrieve other options for each tweet, such as the user name or the geolocation from where the tweet was sent. In this research, the main interest is how messages on Twitter influences the behaviour of (retail) investors, and whether this affects stock price movements. To determine the *virality* of a tweet, the Twitter API allows to retrieve the *public metrics* for each tweet. These *public metrics* are the number of likes, number of retweets, number of quotes, and number of replies associated with a tweet. It shows how many Twitter users interact with a tweet. Hence, these public metrics directly show the number of people that might be moved or influenced by a tweet. Furthermore, these metrics can be used to filter out tweets that are spam or that are not credible, since these tweets do not generate any interactions with other Twitter users.

Exploiting the search query as described above results in the Twitter dataset. This dataset consists of all tweets containing at least one of the words of each row of table 12 in appendix A.1.1, and do not contain any of the words in the company-specific *negation dictionary*. Along with the text, this method also prompts the Twitter API to return for each tweet the user ID of the sender, the time the tweet was created and the number of likes, retweets, quotes and replies (the public metrics).

3.2 Sentiment classification

After retrieving the Twitter data using the Twitter API, the sentiment of the tweets must be determined. This is done using one or more NLP techniques, which are discussed in this section. Here, multiple sentiment classification methods are introduced, and the advantages and disadvantages of these methods are discussed. Furthermore, the data-driven algorithm for selecting the proper method is explained.

3.2.1 Sentiment analysis methods

The goal of sentiment analysis is to correctly *classify* tweets based on their sentiment *polarity* (*i.e.* positive (+1), neutral (0) or negative (-1)) associated with textual data. Classification is the task of sorting data into the corresponding class (assign the data a *label*). Classification methods of textual data can be attributed to either of these types:

- **Supervised classification methods**: Supervised classification methods encompasses all models that make use of a labeled dataset to learn structures and rules which can then be used to assign labels to unseen data. These methods make use of Machine- or Deep Learning approaches and thus require a training set with textual data for which the sentiment is already known (Ribeiro, Araujo, Goncalves, Goncalves, & Benevenuto, 2016). To use supervised classification it is important that a training dataset is retrieved which is very similar to the unseen data. These models are able to derive complex structures, context and negation clauses when properly trained, and are known to perform well in classifying sentiment (Hartmann, Huppertz, Schamp, & Heitmann, 2019).

- **Unsupervised classification methods**: In contrast to supervised classification methods, this type of classification methods do not require a training dataset with labeled items (Ribeiro et al., 2016). In NLP tasks, these methods encompasses all models that make use of *lexicon* based approaches. That means that the classification of tweets is based on a manually constructed sentiment lexicon, in which words or emoticons are assigned a polarity score. By classifying the words in each body of text the total sentiment can be derived. When there is no available labeled dataset these models are known to perform reasonably well in classifying sentiment.

It is clear that under the availability of training data, supervised classification methods perform generally better than unsupervised classification methods, which is supported by the Hartmann et al. (2019). However, when no training data is available, the lexicon based approaches offer reasonable alternatives to classify textual data. Since it is very important to correctly assign the proper polarity label to the tweets in the dataset, the method used to classify the tweets must be properly chosen. Ribeiro et al. (2016) find that the classification performance of the same method on different datasets can be very different, therefore, the chosen sentiment classification model must be able to correctly classify tweets that are similar to the tweets that are used in this research. Before the model selection metrics are presented, this section first discusses two methods that perform generally well in classifying textual data (Go et al., 2009; Ribeiro et al., 2016).

Supervised sentiment classification

Before the sentiment classification model is discussed, the raw Twitter data needs to be *preprocessed*. This appropriate preprocessing steps are different for different NLP tasks, and are therefore finetuned to the training data and model that is chosen. After the preprocessing steps are taken, the textual data is *vectorized*, which is the process of assigning numerical scores to terms in the text (Yang, 1999). The following preprocessing steps modify each individual tweet using common techniques, such that the number of features is reduced (Go et al., 2009), which allows the model to discover links more easily. First, several modifications to the text are made, all letters are set to lower case, all URLs, links, hashtags and username references (@) are removed. Also, all punctuation symbols and numeric digits are removed from the text. Then, all text is *tokenized*, which means that the text is broken into individual words or characters. This process improves the effectiveness of the classification tasks and reduces computational complexity (Yang, 1999).

Subsequently, all elements are assigned a *Part-of-Speech* tag, identifying whether the items are nouns, verbs, adjectives, adverbs, or other parts of speech, which is done because items with the same tag often display similar behaviour within sentences. Lastly, the textual representation is *lemmatized*, this means that words are assigned their base form, as mentioned in the dictionary, with paying attention to the role of these words in the sentence (Yang, 1999). This is done such that the model does not confuse between several words with similar meaning (*e.g.*, walking, walked, walk).

Following these preprocessing steps, the text is vectorized using the Term Frequency - Inverse Document Frequency (*tf-idf*) vectorizer to obtain a numerical representation of the textual data. This vectorizer is commonly used in many NLP approaches (Ramco et al., 2015; Xu, 2018), and assigns each term t a score based on the importance of a certain term within a document (tf) versus the importance of that term across all documents (idf). That is, every term t is assigned a score

$$tf - idf(t, d, D) = tf(t, d) \cdot idf(t, D), \tag{8}$$

where d denotes the document t is in and D denotes the entire corpus, *i.e.* the combined set of documents. Let $f_{t,d}$ denote the number of times term t appears in tweet d, and let n_t denote the number of tweets d in the entire dataset in which the term t is found, then

$$\begin{aligned} tf(t, d) &= \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \\ idf(t, D) &= -\log\left(\frac{n_t}{N}\right), \end{aligned}$$

where N denotes the total number of tweets in the dataset. This method weighs the importance of a term in a single document versus the occurrence of that term in all documents. The terms t from equation (8) are both unigrams (single word) and bigrams (two adjecent words). The tf-idf vectorizer is first fitted to the training dataset, and subsequently the terms t in the unseen data are assigned a score $tf-idf(t,d,D)$ using the weights of the training data. This highlights that the training and unseen data should be familiar, such that the weights are comparable.

After all the tweets in both the training and test dataset are preprocessed and vectorized, the model is trained. Although the Naive Bayes Classifier (NB) performs generally well in multiclass classification assignments (Go et al., 2009; Hartmann et al., 2019), it must be noted that the performance decreases with an increasing number of classes. NB makes use of the posterior probability of a tweet d being assigned to class s, where s denotes the polarity of a tweet (positive, neutral, negative). This probability is characterized by

$$P(s \mid d) = \frac{\left(P(s) \sum_{t \in d} P(tf-idf(t,d,D) \mid s)^{n_t(d)}\right)}{P(d)}, \tag{9}$$

where the probabilities $P(s)$ and $P(tf-idf(t,d,D) \mid s)$ are obtained by maximum likelihood estimates and calculated on the training data. $P(s)$ denotes the probability of an instance being classified as class s, and $P(tf-idf(t,d,D) \mid s)$ denotes the probability of an $tf-idf$ feature occurring in a document labeled as class s. Here $n_t(d)$ is the count of the term frequency - inverse document frequency score $tf-idf(t,d,D)$ in document d. The sentiment polarity class s that maximizes this probability is than assigned to an instance tweet d. Because the NB assumes independence of the features $tf-idf(t,d,D)$, this approach is considered efficient, and easily implemented via the `sklearn` package in Python.

Unsupervised sentiment classification

Unlike supervised classification models, unsupervised models do not require extensive training or a training set. As aforementioned, in NLP tasks, unsupervised classification generally involves the use of a sentiment lexicon, in which words are annotated by a manually determined sentiment score. Ribeiro et al. (2016) investigate in their article multiple of these lexicon based sentiment approaches. The authors conclude that although the performance of these lexicon based approaches varies between the platform and type of text, the VADER (Valence Aware Dictionary for sEntiment Reasoning) classification method performs well on social media documents. This lexicon based approach, constructed by Hutto and Gilbert (2015), is specifically created to handle social media content. Besides its reputable performance on social media texts, an additional benefit of the VADER sentiment lexicon is the fact that it is easily extended. The VADER lexicon handles slang, extensive punctuation (*e.g.* !!?), extended words (*e.g.*, greaaaat), and words in capitals and assigns these different versions of words different sentiment scores. This is particularly useful on Twitter text, where this type of text is very common. Hence, the Twitter data requires very few preprocessing steps, only URLs, hashtags and usernames, and all numeric digits are removed from the text.

The classification using the VADER lexicon is done as follows, in the first place, each word in the text is assigned a valence score, based on the manually created lexicon. Next, the rule based approach deals with negation, relative importance of capitals and punctuation and sentiment shifting clauses (*e.g.* via the word 'but') to derive the compound sentiment score of a specific body of text. Following the approach of Hutto and Gilbert (2015), this compounded score is calculated on every sentence of a tweet, of which the mean is taken to derive the overall sentiment score of a specific tweet, where -1 is the most negative score, and +1 the most positive score a tweet can generate. Similar to the authors, tweets with a score smaller than -0.05 are classified as negative (-1) and tweets with a score above 0.05 are classified as positive (+1), the rest is classified as neutral (0).

3.2.2 Selecting a sentiment classification model

Subsequently, it must be determined how the appropriate sentiment classification model is chosen. After retrieving the Twitter data for each company in this research, a subsample of the data is taken and manually annotated by a panel of eight students. This subsample, containing manually annotated tweets, serves as the test set. The existence of a test set allows to test the proposed models based on actual data used in this research. The model that performs the best on this test set is eventually used to classify the entire Twitter dataset.

However, supervised sentiment classification models require training on a comparable dataset before these models can successfully classify unseen data. This paper proposes to create a training dataset by combining multiple publicly available labeled Twitter datasets. The training dataset is constructed from the publicly available Apple sentiment dataset, a dataset with U.S. airline reviews, a dataset containing preprocessed tweets and lastly a sample of 8000 items from the Sentiment140 dataset[5]. Although this dataset does not contain tweets that are in the exact same domain of the datasets used in this research, they are based on Twitter messages, which might improve the performance of the supervised model. Furthermore, all words that have specific connection to the training data are removed, as the discovered patterns are most likely not applicable on the actual data (*e.g.*, flight or airport have different sentiment meaning in the non-financial Twitter datasets).

To compare the performance of the previously introduced sentiment classification methods, several test statistics can be computed. As Yang (1999) points out, model evaluation should be done on a variety of complementary scores instead of being reliant on a single evaluation score. Therefore, in line with the large-scale sentiment classification comparison done by Ribeiro et al. (2016), both the *Macro* F_1 and the *accuracy* of the models is computed. The model that performs best in classifying the test data on these two metrics is selected in this research. The Macro F_1 is the average of the F_1 score for each class $s \in$ [positive, neutral, negative], which is computed as the harmonic mean between *precision* and *recall* for a class s. Precision $P(s)$ is computed as the number of elements correctly classified as class s, divided by the total number of elements classified as class s. Recall $R(s)$ is the number of elements correctly classified as class s divided by the total number of elements that have true label s (Ribeiro et al., 2016). Let S denote the set with sentiment polarities, and let n_S be the number of elements in S (in this case $n_S = 3$), the Macro F_1 score is then computed by

$$\text{Macro } F_1 = \frac{1}{n_S}\sum_{s\in S} F_1(s) = \frac{1}{n_S}\sum_{s\in S}\frac{2P(s)\cdot R(s)}{P(s)+R(s)}. \tag{10}$$

Along with the Macro F_1 score, the accuracy of the model is calculated as well, to assess which model performs best on the humanly annotated test data. The accuracy is calculated as

$$accuracy = \frac{\#True}{\#True + \#False}, \tag{11}$$

where $\#True$ denotes the number of elements correctly classified, *i.e.* an instance with label s is classified as s, and $\#False$ denotes the number of elements that are falsely classified. This metric shows the fraction of instances correctly classified by a model. The sentiment classification method that achieves the highest rank for these metrics based on the manually annotated test set is chosen in this paper.

Before the metrics are calculated, another method is introduced, which is the Adjusted VADER. As explained, the VADER sentiment lexicon is easily extended, which might prove to be valuable to include

[5]The datasets can be find via these links, Apple dataset, U.S. airline sentiment dataset, tweets dataset, and the Sentiment 140 dataset

domain specific words not included in the regular VADER lexicon (*e.g.* pollution). Recall that sentiment classification methods are very domain sensitive (see Ribeiro et al., 2016), hence domain dependent miss-classifications can be overcome by the introduction of domain words relevant for this research into the VADER lexicon. To determine which words are valuable to consider, the falsely classified tweets from the test dataset by VADER are inspected, and words that signal domain specific characterizations are used to construct the Adjusted VADER lexicon. It must be noted that the addition of items to the lexicon must be done carefully, as resolving every misspecified tweet probably leads to *overfitting*. This is the case when the model is too much tuned according to the training data (in this case the manually annotated test set), and thus fails to correctly classify text that the model has not seen.

3.3 Averaging Twitter sentiment

To get an accurate overview of the sentiment at time t of the entire population active on Twitter, all sentiment scores must be aggregated to reach any conclusion. This section introduces a method to aggregate sentiment polarity across different tweets to derive the sentiment score at time t.

One significant difference between Twitter and more traditional news sources is that Twitter is a crowd-sourced medium (Gupta & Kumaraguru, 2012). This means that any individual can post whatever they want, without being fact-checked before they send their message, which can then influence or inspire other users. As the goal of this research is to model sentiment based on tweets, it is crucial to make a distinction between tweets that are credible and might influence other users (*e.g.* causing them to sell or buy stocks) and tweets that are not credible or contain spam. Since it is important that the derived Twitter-based sentiment score is a good reflection of the actual public sentiment, tweets that are not credible and/or have no influence on other users, should be handled differently than tweets that attract a lot of attention.

Most research that focus on influence and importance of tweets make use of the social structure or the credibility of the sender of the tweet on the basis of the number of followers or historical tweets (Verma, Divya, & Sofat, 2014). Considering the fact that such an approach would require additional information, this paper opts for a *tweet-based*, which makes use of the readily retrieved dataset to tackle this issue.

To determine the sentiment score within a certain interval, previous research does not consider the distinction between more important and less important tweets that are posted in that interval (Fan et al., 2020; Ramco et al., 2015). For example, a tweet sent by a respected investor such as Warren Buffet is weighed equally as a tweet from a consumer expressing its dissatisfaction with a certain product. However, there is a clear difference in the influence of these tweets on the public sentiment and how widespread that sentiment is. Considering that this research explores how Twitter activity affects stock price return volatility, it must be determined to what extent tweets exert influence on Twitter users.

Recall that the Twitter API provides the opportunity to access the public metrics associated with each tweet. Now, these metrics are leveraged to determine the relative importance of each single tweet in the dataset. Since these metrics describe the frequency that Twitter users interact with a tweet, these metrics provide a very good intuition of the relative importance of each tweet and how much attention a tweet generates. In the literature, very little has been written on using these public metrics to find a weighted average of multiple sentiment scores within a certain time period. However, in their article, Perdana and Pinandito (2018) present a method to use like-retweet analysis to assign non-textual scores to retrieved tweets. Consequently, the method to find a public metric based weighted average is based on their findings. The authors calculate a score to determine the importance of each feature $j \in$ [# of likes, # of retweets, # of quotes, # of likes]. To take into account the possible asymmetric effects of the polarity of tweets, this score makes a distinction between positive, negative and neutral polarity. Let j denote public metric j and let S be the set of sentiment polarities (positive, negative, neutral), the score Sc at time t for feature j is defined by

$$Sc_t(j) = \frac{\sum_{s \in S} n_s(\mu_s^j - \mu^j)^2}{(\sigma^j)^2}, \tag{12}$$

where n_s is the number of tweets at t with sentiment polarity s. The average number of feature j for all tweets with polarity s is given by μ_s^j, μ^j denotes the average count of feature j regardless of the polarity and σ^j is the standard deviation of the number of feature j across all tweets. If the standard deviation of the count of feature j across all tweets equals zero, σ^j is set equal to one, this indicates that there is no difference in the counts of public metric j. Equation (12) calculates how much standard deviations the average number of likes, retweets, quotes, or replies for tweets with polarity s deviates from the mean number of likes, retweets, quotes, or replies across all tweets.

The score from equation (12) is calculated at time t using all the tweets send at time τ for $t-1 < \tau \leq t$, which determines the feature score $Sc_t(j)$ at time t for feature j. Using this score, let $m_{k,t}^j$ denote the count of public metric j at time t of a specific tweet k, the nontextual weight for this tweet k at time t is then defined as

$$tweetScore_{k,t} = \left(\sum_{j \in J} m_{k,t}^j \cdot Sc_t(j) \right) + 1, \tag{13}$$

where J denotes the set of all public metrics. In this equation, one is added to account for the case that although $m_{k,t}^j = 0$ for all j, the tweet k is send by a Twitter user, even though it failed to generate any interactions. Unlike Perdana and Pinandito (2018) who combine the sentiment scores and weights, this paper uses the $tweetScore$ to derive a weighted average of all the selected tweets at time t.

The weight for a tweet k from equation (13) is used to determine a weighted average sentiment score that accounts for the relative importance of tweets. This weighted average approach yields the sentiment score at time t using all tweets sent between $t-1$ and t, for $t \in \{1, \dots, T\}$. The weight for tweet k at time t is determined by the normalized tweet weight at time t

$$W_{k,t} = \frac{tweetScore_{k,t}}{\sum_{k=1}^{n_t} tweetScore_{k,t}}, \tag{14}$$

where n_t denotes the number of tweets sent at between $t-1$ and t. Consequently, the Twitter-based sentiment $twitSent$ score at time t is computed by multiplying the sentiment score $sentiScore$ of tweet k by its weight $W_{k,t}$, which yields

$$twitSent_t = \sum_{k=1}^{n_t} W_{k,t} \cdot sentiScore_{k,t}. \tag{15}$$

Hence, equation (15) provides a method to derive the weighted average Twitter derived sentiment at t for $t \in \{1, \dots, T\}$ and can be included in the models that are discussed in more extent later. Here, $sentiScore_{k,t}$ denotes the numerical representation of the sentiment polarity (positive, negative or neutral) as determined by the models from section 3.2. Recall that negative tweets are assigned a score of -1, neutral tweets a score of 0 and positive tweets a score of +1. Thus, negative values of $twitSent_t$ correspond to negative public sentiment at time t, and positive values show positive public sentiment at t, with values closer to zero signalling weaker sentiment and vice versa.

This section introduces a parsimonious method to derive a weighted average sentiment score that takes into account the relative importance of tweets in the conversations happening on Twitter. This method is easily implemented and calculates the relative importance of tweets based on the number of interactions with real-life Twitter users. Therefore, tweets that do not generate any interactions are assigned a very low score, which may help to filter out the importance of spam tweets, and tweets that get a lot of interactions (even though they are spam tweets) are given a higher weight. To this extent, in contrast to other articles, this paper does not immediately discard spam tweets as they might reveal valuable information.

3.4 Twitter attention metrics

In the assessment of Twitter activity on the stock price return volatility, other variables besides the weighted average sentiment must be considered to explore how Twitter influences stock prices. Although the method from equation (15) provides a unique approach to determine the weighted average sentiment score, it does is not a reflection of the total magnitude of the calculated sentiment. To illustrate, ten Twitter users acting very disappointing towards *Coca-Cola* on Twitter between $t-1$ and t results to a high sentiment score $twitSent_t$ in absolute terms. However, 10.000 people tweeting something moderately positive about *Coca-Cola* probably causes an average sentiment score closer to zero, while their feelings are clearly more wide-spread and might have more effect on the stock price volatility of *Coca-Cola*. Therefore, it might turn out to be valuable to consider the total volume of tweets send within each time interval as well. Audrino et al. (2020) find that measures of investors attention variables, specifically the amount of messages posted and the daily financial searches at Google are the most relevant predictors for volatility.

Additionally, as aforementioned in section 3.3 the number of interactions on tweets send between $t-1$ and t present information about the number of people possibly influenced by the tweets. It can be regarded as the number of people agreeing or disagreeing with the content of a tweet, which can be used as a proxy of how much attention a tweet generates. Therefore, the *public metrics* associated with each tweet are included as well to determine how much attention a company receives at time t. As these metrics are already used individually to get a weighted average of the daily sentiment, the metrics associated with all tweets send between $t-1$ and t are summed to derive the daily number of interactions. Denote by $n_{interact,t} = \sum_j^J m_t^j$ the number of interactions at time t, where m_t^j denotes the count of public metric j across all tweets send on day t, $t \in \{1, \dots, T\}$. Hence, the number of interactions helps with assessing the degree of *virality* tweets generated.

To determine the effect of Twitter-based sentiment variables based on non-financial Twitter coverage, several variables are included in this research. In the first place, the sentiment score, weighted by the parsimonious approach introduced in section 3.3 is included. As explained in this section, both the number of tweets and the number of interactions might hold valuable information on the behaviour of the public. Therefore, next to the sentiment, these exogenous variables are included as well to determine how Twitter affects stock price volatility. In line with the literature, it is expected that increased attention, measured by these *attention variables*, signal increased levels of volatility (Audrino et al., 2020). Additionally, it is expected that negative sentiment has a larger effect on stock price return volatility than positive sentiment (M. P. Chen, Chen, & Lee, 2013; Smales, 2015).

4 Modeling volatility

In this chapter, the main econometric models used in this research are explained. These models make use of the previously introduced modeling techniques discussed in section 2. Along with the introduction of the proposed models, this section discusses the associated assumptions needed for consistent estimation, the estimation procedure, and the asymptotic theory of the estimates. Additionally, a varying parameter specification is introduced that is used to determine whether the effect of the Twitter-based sentiment variables is subject to change as a function of the overall state-of-the-economy. Lastly, relevant tests are introduced that allow to test the significance of the Twitter-based variables, and that help to determine the *best-practice* approach of including these variables.

4.1 Econometric models

This section introduces the benchmark specification of the conditional variance model, which will serve as the model against which possible extensions are compared. Subsequently, the proposed methodologies to include exogenous Twitter-based variables into the standard model specification are discussed briefly. Before the estimation procedure of these models is explained, section 4.2 outlines the necessary assumptions and constraints associated with the proposed models.

4.1.1 The benchmark model

Recall from section 2 that the standard GARCH(1,1) process performs generally well in modeling the volatility series of financial time series (Hansen & Lunde, 2005; Lamoureux & Lastrapes, 1990). Consequently, it is assumed that $p = q = 1$ for the models in this research. However, the lags P and Q and the lags including the exogenous variables are allowed to be of a higher order. The exact lag order is later discussed and is determined via a data-driven approach.

The GARCH model formulation deals with the conditional volatility of a time series. However, to impose a drift or moving average structure on the mean of time series, the mean equation needs to be specified as well. Therefore, the conditional mean of the return time series process $r_1, \dots r_T$ for is proposed to follow an ARMA(P, Q) process, that is

$$r_t = \mu + \sum_{i=1}^{P} \gamma_i r_{t-i} + \sum_{i=1}^{Q} \delta_i \varepsilon_{t-i} + \varepsilon_t. \tag{16}$$

Denote by $\vartheta = (\mu, \gamma_1, \dots, \gamma_P, \delta_1, \dots, \delta_Q)'$ the parameter set driving the conditional mean from equation (16), and let the compact parameter space be denoted by $\Theta_0 \in \mathbb{R}^{P+Q+1}$. Subsequently, it is assumed that the residuals from this process, $\varepsilon_1, \dots, \varepsilon_T$ follow a apARCH process specified by equation (1) and equation (6) to accommodate for the *leverage effect*. Hence, the unobservable conditional volatility process σ_t is defined by

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t, \\ \sigma_t^2 &= \omega + \alpha \left(|\varepsilon_{t-1}| - \psi \varepsilon_{t-1}\right)^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{17}$$

where $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\psi \in (-1, 1)$ to guarantee non-negativity. Let $\theta \in \Theta_1$ denote the parameter set containing the parameters of the conditional variance equation, *i.e.*, $\theta = (\omega, \alpha, \beta, \psi)'$. Additionally, let the innovations z_t be an i.i.d. sequence for which $E[z_t \mid \mathcal{F}_{t-1}] = 0$ and $E[z_t^2 \mid \mathcal{F}_{t-1}] = 1$ holds, where $\mathcal{F}_{t-1}$ denote the information until time t, $\{\varepsilon_u, u < t\}$. Recall from section 2 that positive values of ψ indicate that negative shocks of the residuals ε_t have more effect on the conditional variance process of equation (17) than positive shocks, and vice versa.

4.1.2 Exogenous variables

The goal of this paper is to determine whether, and via which methodologies, Twitter-based sentiment variables can be used to predict the conditional volatility of the stock price returns. To this extent, it must be determined how the exogenous variables enter the model defined by equation (16) and equation (17). Section 2 discusses the inclusion of exogenous variables in the conditional variance equation via the method of Han and Kristensen (2014). This research aims to find the best practice approach to include these covariates, hence, three different methodologies are considered that include exogenous variables into the ARMA(P, Q)-apARCH(1,1) model. Let the time series of the exogenous variables be defined by $x_{k,1}, \ldots, x_{k,T}$ where $k \in \{1, \ldots, n_K\}$ is an element of the set of exogenous time series. K is the set of exogenous variables, and n_K the number elements in K. Denote by R_k the number of lagged values of the exogenous variable x_k included in the model. Denote by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ the time series processes of all exogenous variables $k \in \{1, \ldots, n_K\}$, *i.e.*, the vector $\boldsymbol{x}_t$ contains the value at time t of every exogenous variable $k \in \{1, \ldots, n_K\}$.

Model I: ARMAX-apARCH

Before proposing models that incorporate exogenous variables in the conditional volatility, it might prove to be valuable to check how the model performs in terms of likelihood by letting exogenous variables enter the mean equation. That is, the model from equation (16) is rewritten to include n_K lagged exogenous variables $x_{k,t-i}$ with lags $i = 1, \ldots, R_k$. Consequently, define the ARMAX specification by

$$r_t = \mu + \sum_{i=1}^{P} \gamma_i r_{t-i} + \sum_{i=1}^{Q} \delta_i \varepsilon_{t-i} + \sum_{k=1}^{n_K} \left(\sum_{i=1}^{R_k} \pi_{k,i} x_{k,t-i} \right) + \varepsilon_t, \tag{18}$$

where $\mu, \gamma_i, \delta_i, \pi_{k,i} \in \mathbb{R}^{P+Q+1+\sum_{k=1}^{n_K} R_k}$. The residuals from this ARMAX equation follow an apARCH specification, given by equation (17). Thus, this model will be referred to as the ARMAX-apARCH model throughout the remainder of this paper. Denote by $\mathcal{F}_{x,t-1} = \{\varepsilon_u, \boldsymbol{x}_u, u < t\}$ the information set generated by lagged values of ε_t and $x_{k,t}$, then it must hold that $E[z_t \mid \mathcal{F}_{x,t-1}] = 0$ and $E[z_t^2 \mid \mathcal{F}_{x,t-1}] = 1$.

Model II: ARMA-apARCH-apX

To determine the effect of the exogenous variables on the conditional volatility, it must be specified how these variables enter the conditional variance equation described in equation (17). In previously discussed cases, the exogenous covariates entered the GARCH model linearly by squaring their values (see Han and Kristensen, 2014 or Hsu et al., 2021 for examples). However, this specification does not account for possible asymmetric effect of the Twitter-based variables. That is, it is also of interest whether an increase or decrease of the exogenous variables affects the conditional variance σ_t differently, which is supported by Smales (2015). Especially since this thesis includes sentiment scores and *attention variables* (the number of tweets and the number of interactions), it is of interest whether these variables have asymmetric effects on the conditional volatility. Negative sentiment is proposed to have more effect on the volatility of stock price returns than positive sentiment, and similarly high values of attention variables are proposed to have more effect than low values.

To accommodate for these asymmetric effects of the exogenous variables, the exogenous variables enter the model via the function $g(\cdot)$. This function captures possible asymmetric effects of the process $x_{k,t}$ similarly to the asymmetric specification in the apARCH model from equation (17). Define the function $g_k(x_{k,t-i}, \psi_{k,i})$ by

$$g_k(x_{k,t-i}, \psi_{k,i}) = (|x_{k,t-i}| - \psi_{k,i} x_{k,t-i})^2, \tag{19}$$

where $-1 < \psi_{k,i} < 1$ for all $k \in \{1, \ldots, n_K\}$ and $i \in \{1, \ldots, R_k\}$. Recall from section 2 that for negative values of $\psi_{k,i}$ positive values of the i-th order lag of exogenous variable k have more effect on the conditional volatility than negative values, and vice versa. Note that when $\psi_{k,i} \to 1$, the effect of positive values of $x_{k,t-i}$ on the conditional volatility reduces to zero, and similarly, the effect of negative values reduces to zero when $\psi_{k,i} \to -1$. By taking into account possible asymmetry in the effects of the exogenous variables, both the level and the sign of $x_{k,t}$ is accounted for in the effects on the conditional volatility.

To include the asymmetric effects of the exogenous variables, let the function from equation (19) enter the conditional volatility model of equation (17). This results into the following ARMA-apARCH-apX model

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t, \\ \sigma_t^2 &= \omega + \alpha \left(|\varepsilon_{t-1}| - \psi \varepsilon_{t-1} \right)^2 + \beta \sigma_{t-1}^2 + \sum_{k=1}^{n_K} \left(\sum_{i=1}^{R_k} \pi_{k,i} g_k(x_{k,t-i}, \psi_{k,i}) \right), \end{aligned} \tag{20}$$

where ε_t are the residuals from the conditional mean equation equation (16). For the innovations z_t it must hold that $E[z_t \,|\, \mathcal{F}_{x,t-1}] = 0$ and $E[z_t^2 \,|\, \mathcal{F}_{x,t-1}] = 1$, where $\mathcal{F}_{x,t-1}$ is the σ-field generated by $\{\varepsilon_u, \boldsymbol{x}_u, u < t\}$. To ensure non-negativity of the process in equation (20), $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, $\psi \in (-1, 1)$ and with regard to the parameters describing the effect of the exogenous variable, it must hold that $\boldsymbol{\pi}_k \geq 0$ and $-1 < \boldsymbol{\psi}_k < 1$ for $k \in \{1, \ldots, n_K\}$, where $\boldsymbol{\pi}_k = (\pi_{k,1}, \ldots, \pi_{k,R_k})$ and $\boldsymbol{\psi}_k = (\psi_{k,1}, \ldots, \psi_{k,R_k})$ for $k \in \{1, \ldots, n_K\}$. Denote by Θ_1 the entire parameter space containing the parameters of the volatility equation proposed in equation (20), where every realisation satisfies the non-negativity constraints.

In contrast to earlier defined GARCH alternatives that allow for the inclusion of exogenous variables, this method proposes a unique manner which hopefully provides additional insights into the way that these exogenous variables can predict conditional volatility. The asymmetric power function $g(\cdot)$ of equation (19) allows to test the hypotheses that conditional volatility is more affected by negative sentiment, and that high values of the *attention variables* have more effect on the conditional volatility.

Model III: ARMA-apARCH-apXGARCH

In the previous subsections, two rather straightforward approaches were discussed, proposing methodologies to alter the models from equation (16) and equation (17) to account for the effect of exogenous variables. In the article by Fan et al. (2020), the authors find that increased values of a so-called *disagreement measure* has significant effects on the volatility of multiple stock price returns. Therefore, it is proposed that next to the asymmetric effects of the lagged exogenous variables, the degree to which these variables are prone to change can also serve as a relevant predictor of stock price volatility. That is, the model from equation (20) must also account for the degree to which exogenous variables change from day to day. This is comparable with Audrino et al. (2020), who include the standard deviation of the daily sentiment score as a measure of disagreement.

Contrary to Audrino et al. (2020), this research will use the conditional volatility of the exogenous variables as a proxy for the degree of disagreement. That is, this research does not consider the daily disagreement, but considers how the exogenous variables vary over time. To this extent, an ARMA$(1,1)$-GARCH$(1,1)$ structure on the exogenous variables is proposed. This specification is chosen because letting the exogenous variables follow a pure GARCH model seems too restrictive. Hence, let an exogenous variable X_t be defined by the process

$$
\begin{aligned}
X_t &= \mu_x + \gamma_x X_{t-1} + \delta_x \epsilon_{t-1}^{(x)} + \epsilon_t^{(x)} \\
\epsilon_t^{(x)} &= \sigma_t^{(x)} \eta_t^{(x)},
\end{aligned} \tag{21}
$$

where $|\gamma_x|< 1$ must hold for X_t to be stationary and $|\delta_x|< 1$ must hold for invertibility. Let $\vartheta_x = (\mu_x, \gamma_x, \delta_x)'$ denote the parameter set of equation (21). Similar to the earlier introduced GARCH models, without imposing a distribution on the innovations $\eta_t^{(x)}$, it must hold that $\eta_t^{(x)}$ is i.i.d. and $E[\eta_t^{(x)} \,|\, \epsilon_{t-1}^{(x)}, \dots, \epsilon_1^{(x)}] = 0$ and $E[(\eta_t^{(x)})^2 | \epsilon_{t-1}^{(x)}, \dots, \epsilon_1^{(x)}] = 1$. Now, the specification of the conditional variance of the exogenous variable X_t is given by

$$
(\sigma_t^{(x)})^2 = \omega_x + \alpha_x (\epsilon_{t-1}^{(x)})^2 + \beta_x (\sigma_{t-1}^{(x)})^2, \tag{22}
$$

where $\omega_x > 0$, $\alpha_x \geq 0$ and $\beta_x \geq 0$, as the conditional variance process can not be negative. Let the parameter set $\theta_x = (\omega_x, \alpha_x, \beta_x)'$ be a realisation of Θ_x.

Using the specification from equation (22), the degree of variability of $x_{k,t}$ at each time $t \in \{1, \dots, T\}$ is captured by the conditional volatility $(\sigma_t^{(x)})^2$ of the exogenous process. Since the conditional volatility of equation (22) is non-negative and stationary by definition, this process can directly be included into the conditional volatility equation. Thus, the conditional volatility equation from equation (20) is extended to incorporate the conditional volatility process of the exogenous variables as well. Accordingly, this model will be referred to as the ARMA-apARCH-apXGARCH model throughout this research.

To define this model, let equation (16) denote the ARMA(P, Q) process that drives the conditional mean process of the returns r_t. Let K denote the set of exogenous variables that enter the model via the specification in equation (19), and K_1 the set of exogenous variables of which the conditional variance also enters the model, where $K_1 \subseteq K$. Additionally, let R_k denote the lag order for exogenous variable k that enters the model for $k \in \{1, \dots, n_K\}$, and denote by S_k the lag order of the exogenous volatility process that enters the model, where $k \in \{1, \dots, n_{K_1}\}$. Allowing for a more flexible lag structure of the exogenous variables can help pin down how long the effects of shocks to these variables have a lasting effect on the volatility of stock price returns. The ARMA-apARCH-apXGARCH model is defined as

$$
\begin{aligned}
\varepsilon_t &= \sigma_t z_t \\
\sigma_t^2 &= \omega + \alpha \left(|\varepsilon_{t-1}| - \psi \varepsilon_{t-1}\right)^2 + \beta \sigma_{t-1}^2 + \sum_{k=1}^{n_K} \left(\sum_{i=1}^{R_k} \pi_{k,i} g_k(x_{k,t-i}, \psi_{k,i}) \right) + \sum_{k=1}^{n_{K_1}} \left(\sum_{i=1}^{S_k} \lambda_{k,i} \sigma_{t-i}^{(x_k)} \right),
\end{aligned} \tag{23}
$$

where the values of $\sigma_{t-i}^{(x_k)}$ are given by equation (22). To ensure non-negativity of the process described by equation (23), define non-negativity constraints on the parameter space Θ_1. Let $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ and $\psi \in (-1, 1)$, for the parameters describing the effect of the exogenous variables, $\pi_{k,i} \geq 0$, and $\psi_{k,i} \in (-1, 1)$ for $i \in \{1, \dots, R_k\}$, $k \in \{1, \dots, n_K\}$ for the exogenous variables that enter via the function from equation (19), and $\lambda_{k,i} \geq 0$ for $i \in \{1, \dots, S_k\}$, $k \in \{1, \dots, n_{K_1}\}$ for the variables k for which the exogenous volatility process is included.

Lastly, denote by $\mathcal{F}_{x,t-1}$ the information set created by $\{\varepsilon_u, \boldsymbol{x}_u, \boldsymbol{\sigma}_u^{(x)}, u < t\}$, where $\boldsymbol{\sigma}_t^{(x)}$ contains every process $\sigma_t^{(x_k)}$ for $k \in \{1, \dots, n_{K_1}\}$. The innovations z_t must be independent and identically distributed, and it must hold that $E[z_t \mid \mathcal{F}_{x,t-1}] = 0$ and $E[z_t^2 \mid \mathcal{F}_{x,t-1}] = 1$. Other distributional assumptions on the innovations are discussed in the next subsection.

The previous subsection discussed in extent the alternatives of the apARCH model from equation (6), that were briefly touched upon in section 2. Along with the formal model statements, relevant constraints on the parameter space were proposed. These models are used to determine the ways in which exogenous

Twitter-based variables exert influence on the conditional stock price returns. Via these models, it can be tested whether there are observable asymmetric effects of the exogenous variables on the conditional volatility, in line with previously stated propositions. Additionally, equation (23) includes the conditional volatility, such that it can be tested whether the degree of variation of the exogenous variables serves as a relevant predictor for conditional volatility.

4.2 Assumptions and constraints

As discussed in section 2, multiple assumptions and constraints must hold to consistently estimate the QMLE of Bollersev and Wooldridge (1992). In their respective articles, Franq and Zakoian (2004) and Ling and McAleer (2003) show consistency of the QMLE under ARMA-GARCH specifications, and Franq and Thieu (2019) and Han and Kristensen (2014) provide the necessary assumptions for consistency of the QMLE under the GARCH-X and apARCH-X specifications, respectively. The findings in these articles are discussed in this section, and necessary assumptions for the models in section 4.1 are stated. Additionally, this section discusses possible test and evaluation criteria to discuss the stationarity of the exogenous covariates and the optimal lag order for each variable entering the models. This section only discusses relevant assumptions in the case when Twitter-based variables enter the conditional volatility equation, and disregards common identifiability conditions.

Let $\varphi = (\vartheta', \theta')'$ denote the entire set of parameters containing the parameter set of the conditional mean model and the conditional volatility models, respectively. Despite this set being different for each model proposed in section 4.1, this notation will be the same for each different model since the constraints mostly encompass the same parameters and to not confuse the reader with unnecessary super- or subscripts. Let the compact parameter space that satisfies the non-negativity conditions of the conditional volatility models of section 4.1 be denoted by Φ, where $\Phi = \Theta_0 \times \Theta_1$. It is assumed that the parameters set φ is in the interior of the parameter space Φ (which corresponds to Case A and Case C in the article by Franq and Thieu, 2019), this is in line with the assumption from Franq and Zakoian, 2004.

In contrast to the earlier mentioned assumption by Ling and McAleer (2003), the assumption by Franq and Zakoian (2004) does not require the existence of second-order moments to guarantee stationarity of the ARMA process from equation (16) or equation (18). For the ARMA process of the exogenous variables, recall that it must hold that $|\gamma_x| < 1$ and $|\delta_x| < 1$, as this process does not allow for a longer lag structure. Define the AR and MA polynomials by $\mathcal{A}(z) = 1 - \sum_{i=1}^{P} \gamma_i z^i$ and $\mathcal{B}(z) = 1 - \sum_{i=1}^{Q} \delta_i z^i$, to guarantee stationarity of the ARMA process, the following must hold:

Assumption 1 $\mathcal{A}(z)\mathcal{B}(z) = 0$ *implies that* $|z| > 0$.

Note that for the residuals ε_t of the conditional mean model, it holds that $\varepsilon_t = \sigma_t z_t$. Then, following the assumption by Franq and Thieu (2019), the following must hold for the innovations of the conditional volatility model of both the conditional volatility models of the return, as well as for the innovations $\eta_t^{(x)}$ from equation (21).

Assumption 2 $(z_t, \eta_t^{(x)}, \boldsymbol{x}_t)$ *is a strictly stationary process, and there exists* $s > 0$ *such that* $E|z_1|^s < \infty$, $E|\eta_1^{(x)}|^s < \infty$ *and* $E||\boldsymbol{x}_1||^s < \infty$.

As the solution to equation (22) must also satisfies the assumptions, the process $\boldsymbol{\sigma}_t^{(x)}$ is strictly stationary by definition. For the returns to be a strictly stationary process described by the models in equation (17), the following must hold as well:

Assumption 3 z_t *and* $\eta_t^{(x)}$ *are i.i.d. sequences for* $t \in \{1, \dots, T\}$, *and* $E[z_t|\mathcal{F}_{x,t-1}] = 0$ *and* $E[z_t^2|\mathcal{F}_{x,t-1}] = 1$ *holds for* z_t, *and for* $\eta_t^{(x)}$, $E[\eta_t^{(x)} \mid \epsilon_{t-1}^{(x)}, \dots, \epsilon_1^{(x)}] = 0$ *and* $E[(\eta_t^{(x)})^2 \mid \epsilon_{t-1}^{(x)}, \dots, \epsilon_1^{(x)}] = 1$.

Assumption 4 $l = \alpha(1+\psi^2)+\beta < 1$ *for all* $\varphi \in \Phi$, *and* $\beta_x + \alpha_x < 1$ *for all* $\theta_x \in \Theta_x$.

Recall that for model I and model II, the σ-field containing all relevant information up to time t was specified by $\mathcal{F}_{x,t-1} = \{\varepsilon_u, \boldsymbol{x}_u, u < t\}$, and in the case for model III $\mathcal{F}_{x,t-1} = \{\varepsilon_u, \boldsymbol{x}_u, \boldsymbol{\sigma}_u^{(x)}, u < t\}$. Under these assumptions, following Franq and Thieu (2019) and Franq and Zakoian (2004) the return series is a strictly stationary solution of the models proposed in section 4.1. Note that assumption 4 guarantees stationarity of the conditional volatility process, and l captures the degree to which shocks to the conditional variance equation are persistent.

Other assumptions mentioned by the authors consider multicollinearity in the exogenous variables, and the incorporation of redundant exogenous variables (*i.e.*, Franq and Thieu (2019) give a remark about letting lagged residual values ε_t enter as exogenous variables), which are generally assumed hold throughout this paper and do not need additional specification.

Checking stationarity

From assumption 2, it can be seen that the exogenous processes must satisfy stationarity. A stationary process is a process of which the statistical attributes, mean and standard deviation, do not change over time. In most cases, time series experiencing a trend over time are non-stationary and hence prohibit the QMLE from being consistent. To this extent, the Augmented Dickey-Fuller (ADF) test is employed (Dickey & Fuller, 1979). Via this test it is tested whether a *unit root* is present in a time series process. Let Y_t be a time series process and denote by $\Delta Y_t = Y_t - Y_{t-1}$, then the following function is considered

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \delta_1 \Delta Y_{t-1} + \ldots + \delta_p \Delta Y_{t-p} + \varepsilon_t.$$

Dickey and Fuller (1979) propose a test for the null hypothesis that $\rho = 0$, which is equivalent to testing for the existence of unit roots. The process Y_t is stationary if the test statistic $DF = \frac{\hat{\rho}}{s.e(\hat{\rho})}$ exceeds the critical value at the $p\%$ significance level, which rejects the null at $p\%$ significance. The values for DF can be found in Fuller (1976).

By applying this test on the exogenous variables $x_{k,t}$ it can be tested whether these processes are stationary and hence if assumption 4 is satisfied. If for any $k \in \{1, \ldots, n_K\}$ the null hypothesis can not be rejected, differencing is applied to this variable and the process $\Delta x_{k,t} = \log(x_{k,t} - x_{k,t-1})$ is again checked for unit roots and enters the model if the ADF test is rejected. If again the null hypothesis can not be rejected, this variable is omitted from this research, as iteratively taking the logarithmic difference of a variable will lead to un-interpretable parameter estimates.

Variable lags

Before the QMLE can be found, the optimal number of lags in the model must be determined. The most popular approaches to test for the optimal level of model complexity are the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) (Schwarz, 1978). Both methods present a weigh-off between model complexity and maximum likelihood, with the latter punishing the inclusion of more variables more.

To select the optimal number of lags in both the ARMA(P, Q) conditional mean model as well as the optimal number of lags of the Twitter-based variables, the BIC is employed. As model complexity unnecessarily rises when evaluating all possible combinations of P, Q and all values R_k and S_k, selecting the lags in the conditional mean model and the lags for each exogenous variable is done separately. Define by $\mathcal{L}_T$ be the quasi log-likelihood of a model with n parameters and sample size T, then the BIC is defined by

$$BIC = n \log T - 2 \cdot \mathcal{L}_T. \tag{24}$$

The AR and MA lags P and Q that minimize this function, using the specification by equation (16) and equation (17) is applied to all models that allow for the inclusion of exogenous covariates. The lag order of the exogenous variables, R_k and S_k is determined by the values that minimize equation (24).

4.3 Parameter estimation

Recall from section 2 that the most widely used method to estimate the parameters of GARCH alternatives is by maximizing the Quasi Maximum Likelihood. Under the assumptions from section 4.2, the Quasi Maximum Likelihood estimator provides consistent estimates of the true parameter values $\varphi = (\vartheta', \theta')'$. Asymptotic properties and consistency of the parameters are remained if the QMLE specification is not oversimplified (Bollersev & Wooldridge, 1992). This section provides the quasi maximum likelihood function and the proper estimation procedure, laid out by Franq and Thieu (2019). Additionally, the asymptotic properties and consistency of the QMLE are shown.

4.3.1 Quasi Maximum Likelihood Estimator

To determine the QMLE for the period $t \in \{1, \ldots, T\}$, start by initializing lagged values for the conditional mean and conditional variance equation $\tilde{\varepsilon}_{1-Q}, \ldots \tilde{\varepsilon}_0$, $r_{1-P}, \ldots, r_0$, $\tilde{\sigma}_0 \geq 0$, and the stationary exogenous variables $x_{k,1-R_k}, \ldots, x_{k,0}$ for $k \in \{1, \ldots, n_K\}$. Following Franq and Zakoian (2004), the ARMA(P, Q) residuals for $t \in \{1, \ldots, T\}$ are defined by

$$\tilde{\varepsilon}_t(\vartheta) = \begin{cases} r_t - \mu - \sum_{i=1}^{P} \gamma_i r_{t-i} - \sum_{i=1}^{Q} \delta_i \tilde{\varepsilon}_{t-i}, \\ r_t - \mu - \sum_{i=1}^{P} \gamma_i r_{t-i} - \sum_{i=1}^{Q} \delta_i \tilde{\varepsilon}_{t-i} - \sum_{k=1}^{n_K} \left(\sum_{i=1}^{R_k} \pi_{k,i} x_{k,t-i} \right), \end{cases} \tag{25}$$

where the conditional mean models follows equation (18) or equation (16), respectively. Note that the first case also corresponds to the benchmark model from equation (17). For the parameters in this model, assumption 1 must hold, as does assumption 2 for the exogenous variables.

Recall from section 4.1 that model III included the conditional volatility of the exogenous variables in K_1 via the function from equation (22). Let $\tilde{\boldsymbol{\epsilon}}_t^{(x)}$ and $\tilde{\boldsymbol{\sigma}}_t^{(x)}$ denote the residuals and the conditional volatility of these exogenous variables at time t, respectively. Denote $\boldsymbol{x}_t^{(sub)}$ denote the subset of exogenous variables contained in K_1, such that $\boldsymbol{x}_t^{(sub)} \subseteq \boldsymbol{x}_t$. Recall that to initialize the process in equation (25) the lagged exogenous variables $x_{k,1-R_k}, \ldots, x_{k,0}$ were defined. Hence, if $S_k > R_k$ for $k = 1, \ldots n_{K_1}$, define $x_{k,-S_k}^{(sub)}, \ldots, x_{k,-R_k-1}^{(sub)}$. Next, define for each variable in K_1 initial values of $\tilde{\epsilon}_{-S_k}^{(x)}$ and $\tilde{\sigma}_{-S_k}^{(x)}$. The ARMA-GARCH process of the exogenous variables for $t \in \{1 - S_k, \ldots, T\}$ is defined by

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}}_t^{(x)}(\boldsymbol{\vartheta}_x) &= \boldsymbol{x}_t^{(sub)} - \boldsymbol{\mu}_x' - \boldsymbol{\gamma}_x' \boldsymbol{x}_{t-1}^{(sub)} - \boldsymbol{\delta}_x' \tilde{\boldsymbol{\epsilon}}_{t-1}^{(x)} \\ \tilde{\boldsymbol{\sigma}}_t^{(x)}(\boldsymbol{\theta}_x) &= \sqrt{\boldsymbol{\omega}_x' + \boldsymbol{\alpha}_x' (\boldsymbol{\epsilon}_{t-1}^{(x)})^2 + \boldsymbol{\beta}_x' (\boldsymbol{\sigma}_{t-1}^{(x)})^2}, \end{aligned} \tag{26}$$

where $\boldsymbol{\mu}_x = \left(\mu_{x,1}, \ldots, \mu_{x,n_{K_1}} \right)'$ and the same holds for all other parameters in $\boldsymbol{\vartheta}_x$ and $\boldsymbol{\theta}_x$. Recall that the innovations form an unobserved process, and are defined by equation (21). This specification ensures existence of the S_k-th order lag of the estimated exogenous volatility at time $t = 1$, although the notation may be slightly confusing.

As the estimates for the exogenous volatility process are defined by the method above, the conditional volatility via the methods proposed in section 4.1 can be constructed. To define the conditional volatility of either of the three models, start by letting $\tilde{\boldsymbol{\sigma}}_t^{(x)} = \left(\tilde{\sigma}_t^{(x_1)}, \ldots, \tilde{\sigma}_t^{(x_{n_{K_1}})} \right)'$ which was defined previously.

Then, the conditional volatility is specified by

$$\begin{aligned}
\tilde{\sigma}_t^2(\theta) &= \omega + \alpha\left(|\tilde{\varepsilon}_{t-1}| - \psi\tilde{\varepsilon}_{t-1}\right)^2 + \beta\tilde{\sigma}_{t-1}^2 \\
\tilde{\sigma}_t^2(\theta) &= \omega + \alpha\left(|\tilde{\varepsilon}_{t-1}| - \psi\tilde{\varepsilon}_{t-1}\right)^2 + \beta\tilde{\sigma}_{t-1}^2 + \sum_{k=1}^{n_K}\left(\sum_{i=1}^{R_k}\pi_{k,i}g_k(x_{k,t-i},\psi_{k,i})\right) \\
\tilde{\sigma}_t^2(\theta) &= \omega + \alpha\left(|\tilde{\varepsilon}_{t-1}| - \psi\tilde{\varepsilon}_{t-1}\right)^2 + \beta\tilde{\sigma}_{t-1}^2 + \sum_{k=1}^{n_K}\left(\sum_{i=1}^{R_k}\pi_{k,i}g_k(x_{k,t-i},\psi_{k,i})\right) + \sum_{k=1}^{n_{K_1}}\left(\sum_{i=1}^{S_k}\lambda_{k,i}\tilde{\sigma}_{t-i}^{(x_k)}\right),
\end{aligned} \tag{27}$$

where the functions correspond to the conditional volatility described by the models from equation (17), equation (20) and equation (23), respectively. To ensure non-negativity, it must hold that $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\psi \in (-1, 1)$, and these parameters must satisfy assumption 4 as well. For the covariates, $\pi_{k,i} \geq 0$ for $i \in \{1, \ldots, R_k\}$, $k \in \{1, \ldots, n_K\}$, $\lambda_{k,i} \geq 0$ for $i \in \{1, \ldots, S_k\}$, $k \in \{1, \ldots, n_{K_1}\}$, and $\psi_{k,i} \in (-1, 1)$ for $i \in \{1, \ldots, R_k\}$, $k \in \{1, \ldots, n_K\}$. Again, the innovations z_t follow equation (1), and must satisfy assumption 2 and assumption 3.

The ARMA residuals from equation (25) and the conditional volatility as defined above gives values of $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_T$ and $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_T$. While Bollersev and Wooldridge (1992) provide the conditions for consistency of the Gaussian quasi log-likelihood, Franq and Thieu (2019) use the following negative quasi log-likelihood function

$$\ell_t(\varphi) = \log\sigma_t(\theta) + \frac{\varepsilon_t^2(\vartheta)}{\sigma_t^2(\theta)}.$$

To find the parameter estimates $\hat{\varphi} = (\hat{\vartheta}', \hat{\theta}')'$ that are consistent estimates of the true parameter values in φ, define the QMLE as the argument that minimizes

$$\mathcal{Q}_T(\varphi) = \frac{1}{T}\sum_{t=1}^{T}\tilde{\ell}_t(\varphi) = \frac{1}{T}\sum_{t=1}^{T}\left(\log\tilde{\sigma}_t(\theta) + \frac{\tilde{\varepsilon}_t^2(\vartheta)}{\tilde{\sigma}_t^2(\theta)}\right), \tag{28}$$

which is equivalent to finding the parameters that maximize the quasi log likelihood function. Thus, the QMLE is defined as any measurable solution $\hat{\varphi}$ that minimizes $\mathcal{Q}_T(\varphi)$, *i.e.*, $\hat{\varphi} = \underset{\varphi\in\Phi}{\operatorname{argmin}}\, \mathcal{Q}_T(\varphi)$.

Under the assumptions and constraints explained in section 4.2, Franq and Thieu (2019) prove that when $T \to \infty$ then the solution of equation (28) converges in probability to the actual parameter values φ, *i.e.* $\hat{\varphi} \to \varphi$ which means consistency of the QMLE.

Minimization of $\mathcal{Q}_T(\varphi)$ yields the same estimates as maximization of the quasi log-likelihood $-\mathcal{Q}_T(\varphi)$. Since multiple metrics and tests (see for instance the BIC statistic in equation (24)) consider a maximized (quasi) log-likelihood value, throughout this paper the maximum quasi log-likelihood is defined by $\mathcal{L}_T(\hat{\varphi}) = -\mathcal{Q}_T(\hat{\varphi})$, under the parameters estimates that minimize equation (28).

Note that in order to estimate the parameters of the ARMA-apARCH-apXGARCH model from equation (23), the conditional volatility process of the exogenous variables $\boldsymbol{x}^{(sub)}$ must be determined. As the exogenous variables were proposed to also follow an ARMA-GARCH process equation (28) can be used in a similar fashion to find the optimal parameters $\tilde{\boldsymbol{\vartheta}}_x$ and $\tilde{\boldsymbol{\theta}}_x$. That is, the QMLE $\tilde{\boldsymbol{\varphi}}_x$ which form the exogenous ARMA-GARCH process is found by

$$\tilde{\boldsymbol{\varphi}}_x = \underset{\varphi\in\Phi}{\operatorname{argmin}}\frac{1}{T}\sum_{t=1}^{T}\left(\log\tilde{\boldsymbol{\sigma}}_t^{(x)}(\boldsymbol{\theta}_x) + \frac{(\tilde{\boldsymbol{\epsilon}}_t^{(x)}(\boldsymbol{\vartheta}_x))^2}{(\tilde{\boldsymbol{\sigma}}_t^{(x)}(\boldsymbol{\theta}_x))^2}\right), \tag{29}$$

which is a consistent estimate of $\boldsymbol{\varphi}_x = \left(\varphi_{x,1}, \ldots, \varphi_{x,n_{K_1}}\right)'$ under the assumptions from section 4.2. Subsequently, these parameter estimates are used to construct the conditional volatility process of equation (26) of exogenous variable $k = 1, \ldots, n_{K_1}$ for $t \in \{1-S_k, \ldots, T\}$ which enters the ARMA-apARCH-apXGARCH model.

4.3.2 Asymptotic properties of the QMLE

Franq and Thieu (2019) provide the asymptotic properties of the QMLE by defining four distinct cases, as the asymptotic behaviour is different when one or more values in $\hat{\varphi}$ are zero, or the residuals $\tilde{\varepsilon}_t$ and the exogenous variables are not independent of the innovations z_t. Note that section 4.2 already stated that only parameter values that do not stand at the boundary are considered. Additionally, it is assumed that the innovations z_t are independent of the information set described by $\mathcal{F}_{x,t-1}$ (which corresponds to case A in the article by Franq and Thieu, 2019).

To define the asymptotic distribution of the QMLE, an additional assumption must be made on the fourth moment of the innovations z_t,

Assumption 5 $E[z_t^4] < \infty$ *for the conditional volatility models, and* $E[(\eta_t^{(x)})^4] < \infty$ *for the model in equation* (22).

Assumption 5 guarantees the existence of the variance of the score function $s_t(\varphi) = \frac{\partial \ell_t(\varphi)}{\partial \varphi}$. Under the assumptions from section 4.2 and assumption 5, the QMLE $\hat{\varphi}$ is a consistent estimator of the true parameter values φ. Franq and Thieu (2019) define under the limiting distribution of the QMLE by

$$\sqrt{T}(\hat{\varphi} - \varphi) \sim N(0, \Sigma), \qquad \text{where } \Sigma = \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}. \tag{30}$$

Hence, the distribution of the QMLE is asymptotically normal with variance-covariance matrix given by Σ. Combining the asymptotic results in Franq and Zakoian (2004) of the ARMA-GARCH model and from Franq and Thieu (2019) who show the asympotics of the apARCH-X model, the Hessian $\mathcal{J}$ is defined as follows:

$$\mathcal{J} = E\left[\frac{\partial^2 \ell_t(\varphi)}{\partial \varphi \partial \varphi'}\right] = E\left[\frac{1}{\sigma_t^4(\theta)} \frac{\partial \sigma_t^2(\theta)}{\partial \varphi} \frac{\partial \sigma_t^2(\theta)}{\partial \varphi'}\right] + 2E\left[\frac{1}{\sigma_t^2(\theta)} \frac{\partial \varepsilon_t(\vartheta)}{\partial \varphi} \frac{\partial \varepsilon_t(\vartheta)}{\partial \varphi'}\right].$$

Following Franq and Zakoian (2004), $\mathcal{I}$ is defined as

$$\mathcal{I} = E\left[\frac{\partial \ell_t(\varphi)}{\partial \varphi} \frac{\partial \ell_t(\varphi)}{\partial \varphi'}\right].$$

Franq and Thieu (2019) propose in proposition I in their article that under equation (30), strongly consistent estimators of $\mathcal{J}$ and $\mathcal{I}$ can be defined as follows

$$\begin{aligned} \hat{\mathcal{J}}_T &= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\tilde{\sigma}_t^4(\hat{\theta})} \frac{\partial \tilde{\sigma}_t^2(\hat{\theta})}{\partial \varphi} \frac{\partial \tilde{\sigma}_t^2(\hat{\theta})}{\partial \varphi'} + \frac{2}{T}\sum_{t=1}^{T} \frac{1}{\tilde{\sigma}_t^2(\hat{\theta})} \frac{\partial \tilde{\varepsilon}_t(\hat{\vartheta})}{\partial \varphi} \frac{\partial \tilde{\varepsilon}_t(\hat{\vartheta})}{\partial \varphi'}, \\ \hat{\mathcal{I}}_T &= \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \tilde{\ell}_t(\hat{\varphi})}{\partial \varphi} \frac{\partial \tilde{\ell}_t(\hat{\varphi})}{\partial \varphi'}. \end{aligned}$$

Combining the equations for the Hessian and the score of the QMLE, the variance-covariance matrix Σ can be estimated by[6]

$$\hat{\Sigma}_T = \hat{\mathcal{J}}_T^{-1} \hat{\mathcal{I}}_T \hat{\mathcal{J}}_T^{-1}. \tag{31}$$

Note that these equations also hold for the asymptotic distribution of the parameters $\tilde{\varphi}_{x,k}$ of the exogenous conditional volatility equation, given by equation (21) and equation (22) under the assumptions from section 4.2 and assumption 5.

[6]Calculations of the Hessian and score function are done by a finite-difference approach in Python.

4.4 Varying parameter specification

In general GARCH-type models, a linear relationship between the lagged residuals, AR part and possible exogenous variables is commonly assumed to model the conditional volatility process. However, it can be argued that the dynamics of the volatility process are prone to change through time. Andreou and Ghysels (2002) test the existence of structural breaks in the volatility dynamics and find alignment with Asian and Russian financial crises. Bauwens, Preminger, and Rombouts (2010) employ a Markov switching regime GARCH model to account for these structural breaks, and conclude this model provides a better fit on S&P500 volatility than a single regime ARCH model. Additionally, both Kim and Kon (1999) and Lamoureux and Lastrapes (1990) provide evidence that a model that is not able to capture deterministic regime-shifts is likely to overstate the degree of persistence.

As this research is concerned with investigating the role of non-financial public sentiment on stock price volatility, it might be valuable to see whether the effect of Twitter-based variables is different during different states of the economy. Amado and Teräsvirta (2008) investigate the use of parametric time dependent extensions of the standard GARCH model, where the parameters are allowed to change as functions of time. The authors propose a modification of the parameters to support a smooth transition over time. That is, via the additive structure which they propose, the parameter vector θ is transformed into the time varying vector

$$\theta(t) = \theta + \theta^* G(t, \nu, c),$$

where the continuous function $G(t, \nu, c)$ takes into account the smooth transition between parameter values and takes values between zero and one. However, in contrast to Amado and Teräsvirta (2008) and in line with the master thesis of Thomassen (2018), this research does not focus on a smooth transition through time, but rather on stages of the economy, *i.e.* whether at time t the economy is in a downturn/investors are very *bearish*. Let ζ_t denote an exogenous variable describing the state of the economy (*e.g.*, ζ_t contains macroeconomic variables such as housing starts or GDP). The choice of ζ_t used in this research is specified later. Then, similar to Thomassen (2018), define the state-of-the-economy dependent parameter vector $\theta(\zeta_t)$ by

$$\theta(\zeta_t) = \theta + \theta^* G(\zeta_t, \nu, \bar{\zeta}), \tag{32}$$

where $\bar{\zeta}$ denotes the sample mean of ζ_t for $t \in \{1, \ldots, T\}$. In the article by Amado and Teräsvirta (2008), c does not necessarily have to be the sample mean, it can be defined as the point where a structural break occurs. However, for simplicity only deviations from the sample mean are defined as structural breaking points. Amado and Teräsvirta (2008) define the transition function $G(\zeta_t, \nu, \bar{\zeta})$ by

$$G^*(\zeta_t, \nu, \bar{\zeta}) = \left(1 + \exp\{-\nu(\zeta_t - \bar{\zeta})\}\right)^{-1}, \tag{33}$$

where $\nu > 0$, which is the general logistic function. Note that for values of ζ_t below the sample mean $\bar{\zeta}$ this function goes to one, and when ζ_t exceeds the sample mean, it approaches zero for increasing values of ζ_t. In equation (33), the parameter ν defines the level of smoothness in the transition function, it determines how smooth the parameter values fluctuate between θ and $\theta + \theta^*$. More specifically, for low values of ν this transition is rather smooth, and when $\nu \to \infty$, the transition from one parameter vector to the next is abrupt, approaching a structural break.

Taking the function of Amado and Teräsvirta (2008) that allows for multiple structural breaks as a starting point, the following is proposed. Let $\boldsymbol{\zeta}_t = (\zeta_{1,t}, \ldots, \zeta_{n,t})$ be a set of n exogenous variables that might provide useful in detecting time-varying regimes. Then the smooth logistic transition function indicating the appropriate state is given by

$$G(\boldsymbol{\zeta}_t, \nu, \bar{\boldsymbol{\zeta}}) = \left(1 + \exp\left\{ -\nu\left(\frac{1}{n}\sum_{i=1}^{n}(\zeta_{i,t} - \bar{\zeta}_i)\right)\right\}\right)^{-1}, \tag{34}$$

which provides a generalization of equation (33). This function has the benefit that it can include multiple exogenous variables. Structural breaks are only observed if across all items in $\boldsymbol{\zeta}_t$, the deviation from the sample means $\boldsymbol{\zeta}$ is negative. As equation (34) allows for the inclusion of multiple exogenous variables, multiply proxies can be used to account for switching states of the economy. This reduces dependence on a single metric that is said to indicate the state of the economy, and only indicates a break when the items in $\boldsymbol{\zeta}_t$ mutually deviate from their sample means.

Define the models that allow for time-varying parameters by replacing the parameter vector θ by $\theta(\zeta_t) = \theta + \theta^* G(\zeta_t, \nu, \bar{\zeta})$ in the models from equation (27). In order to not drastically increase the number of parameters, the goal of this varying parameter specification is to check whether the conditional volatility can be better characterized by allowing for regime-depending parameters, therefore, only the parameters in θ are altered to accommodate for time varying effects. Then, define the total parameter set $\varphi(\zeta_t) = (\vartheta', \theta(\zeta_t))'$, which can be estimated similarly to the models with a constant parameter specification. That is, the parameters $\varphi(\zeta_t)$ that minimize $\mathcal{Q}_T$ from equation (28) are the parameters that define the conditional volatility that accommodates for possible regime shifts. The assumptions on the parameter space stated in section 4.2 must hold for every value of $\varphi(\zeta_t)$ at time t.

To determine the optimal value of the smooth transition operator ν, a grid search on a limited number of options is performed to provide an estimate for the range where ν is optimal. Thomassen (2018) reports that including ν in the parameter set leads to high standard errors of the parameter estimates, caused by the fact that the quasi log likelihood is insensitive to small changes in ν. By conducting a grid search on a relatively low number of options, the behaviour of the quasi log-likelihood as a function of ν is investigated to provide an intuition for the optimal value of the smooth transition operator. This method is discussed in more extent in section 5.

By accommodating for varying parameters of the conditional volatility equation using the function equation (34), the relationship between Twitter activity and stock price return volatility can be investigated more in-depth. The specification allows to distinguish the effect of Twitter through various stages of the economy, which could bring interesting insights not yet observed. For instance, positive values of π_k^* indicate that the effect of exogenous variable k is greater in times when $G(\cdot) = 1$, and contrarily, when $\pi_k < \pi_k^* < 0$ the effects of the exogenous variable k on the conditional volatility reduces when $G(\cdot) = 1$. Additionally, from Kim and Kon (1999) and Lamoureux and Lastrapes (1990) it is expected that the total degree of persistence $l = \alpha(1 + \psi^2) + \beta$ decreases when allowing for varying parameters.

4.5 Statistical tests

To determine whether any of the apARCH models introduced in equation (20) and equation (23) explains the volatility better than the benchmark model from equation (17), multiple tests are conducted. First, the significance of individual coefficients are tested. This helps to identify via which way, if any, Twitter-based variables exert influence on the conditional volatility process. Additionally, two algorithms are introduced to compare the in-sample fit of various models with each other which will help in determining the *best-practice* approach to include Twitter-based variables.

The parameters of the models in section 4.4 can be tested similarly to the models from section 4.1. The tests in this section additionally allow to test the in-sample fit of the varying-parameter models against the models that assume a constant parameter specification.

4.5.1 Significance of parameters

In the first place, the significance of each element in $\hat{\varphi}$ can be tested via the popular t-test. This test is easily employed and tests whether individual elements k in $\hat{\varphi}$ are significantly different from zero, *i.e.*,

$$H_0 : \hat{\varphi}_k = 0 \text{ and } \hat{\varphi}_{j \neq k} \neq 0 \qquad \text{vs.} \qquad H_a : \hat{\varphi} \neq 0,$$

which tests for all individual parameter estimates in $\hat{\varphi}$ if they differ from zero. When the assumptions stated in section 4.2 hold, both Franq and Thieu (2019) and Han and Kristensen (2014) show that the t-test statistic to test the nullity of element k in $\hat{\varphi}$ is defined by

$$t_k = \frac{\hat{\varphi}_k}{s.e.(\hat{\varphi}_k)}, \tag{35}$$

where $s.e.(\hat{\varphi}_k) = \text{diag }(\hat{\Sigma}_T)_k^{\frac{1}{2}}$, which corresponds to the square root of the k-th element on the diagonal of the variance-covariance matrix defined in equation (31). H_0 is rejected at significance level α when $t_k > \Phi^{-1}(1 - 2\alpha)$, where Φ denotes the standard normal cumulative distribution function. When H_0 can not be rejected at the significance level α, this indicates that the k-th element of $\hat{\varphi}$ is not significantly different from zero, and thus does not have any predictive power on the conditional volatility.

4.5.2 Comparison of models

To test whether the advanved apARCH-X type models increases the model performance in terms of the quasi log likelihood $\mathcal{L}_T$, the likelihood ratio test is employed. Under the assumptions and constrains stated in section 4.2, Franq and Zakoian (2009) propose the use of the Quasi Likelihood Ratio (QLR) test statistic to test whether a more advanced model provides any benefits over a *restricted* model. Under this restriction, several parameters are kept equal to zero. This allows for the comparison of the proposed models from section 4.1 and their varying-parameter counterparts with the benchmark model from equation (17). This test can additionally be employed to compare the constant parameter models from section 4.1 with the models that allow for varying-parameters via the function from equation (34). Lastly, since the ARMA-apARCH-aX model is nested in the ARMA-apARCH-apXGARCH model, this test allows to test whether inclusion of the so-called *disagreement measures* $\sigma_t^{(x_k)}$ can lead to significant model improvement.

Let $\varphi^{(1)}$ and $\varphi^{(2)}$ be two components of the parameter vector φ and let m_i for $i = 1, 2$ denote the number of elements in component 1 and 2, respectively. To compare the performance of several models against each other, the following hypothesis is tested

$$H_0 : \varphi^{(2)} = \mathbf{0}_{m_2} \qquad \text{vs.} \qquad H_a : \varphi^{(2)} \neq \mathbf{0}_{m_2},$$

under the assumption that all elements in $\varphi^{(1)}$ are strictly nonzero. Let φ_0 denote the parameter vector under H_0, so with $\varphi^{(2)} = \mathbf{0}_{m_2}$ and let φ denote the parameter vector under the alternative. To use this test, two models must be nested, in order to constrain several parameters to be equal to zero, that is, φ_0 and φ must contain the same parameters.

Following Franq and Zakoian (2009), the Quasi Likelihood Ratio statistic Λ_{QLR} is given by

$$\Lambda_{QLR} = -2\left(\mathcal{L}_T(\hat{\varphi}_0) - \mathcal{L}_T(\hat{\varphi})\right), \tag{36}$$

where $\hat{\varphi}_0$ and $\hat{\varphi}$ are the parameter values that minimize $\mathcal{Q}_T$ under the null and alternative hypothesis, respectively. The null hypothesis is rejected for large values of the QLR, following Wilks (1938), Λ_{QLR} follows a $\chi^2_{m_2}$ distribution. Note that the degrees of freedom correspond to the number of elements in $\varphi^{(2)}$ assumed to be equal to zero. Hence, if the null hypothesis can not be rejected at significance level α, it can be concluded that a more complex model does not lead to significant improvement of the Quasi Likelihood of the conditional volatility model.

4.5.3 Evaluation of the in-sample fit

Recall that the conditional volatility process is an unobserved process, and hence no straightforward techniques (such as the RMSE) can be applied to evaluate the in-sample fit or forecasting performance of a model. However, in conditional volatility modeling, it is common to use a conditionally unbiased estimator as a *proxy* for the true unobserved conditional volatility process (Patton, 2011). The author considers multiple loss functions and ranks the forecasts based on these function, with the goal to define the loss functions that are 'robust' to imperfect volatility proxies. Denote by $\hat{\sigma}^2_{t,prox}$ the proxy of the conditional volatility, which is estimated by $\tilde{\sigma}^2_t$ from equation (27). Hansen and Lunde (2006) point out that the goal of ranking multiple volatility in-sample fits is a qualitative task rather than a quantitative one, especially as an imperfect volatility proxy is chosen to rank the in-sample fit of the model. Following Patton (2011), two robust loss functions are proposed to evaluate the in-sample fit of the volatility models using the proxy $\hat{\sigma}^2_{t,prox}$, which are the Mean Squared Error (MSE) and the Quasi Likelihood (QLIKE) loss functions. The errors of a volatility series using these methods is given by

$$\begin{aligned} L(\hat{\sigma}^2_{prox}, \tilde{\sigma}^2) &= \frac{1}{T}\sum_{t=1}^{T}(\hat{\sigma}^2_{t,prox} - \tilde{\sigma}^2_t)^2, & \text{(MSE)} \\ L(\hat{\sigma}^2_{prox}, \tilde{\sigma}^2) &= \frac{1}{T}\sum_{t=1}^{T}\left(\log(\tilde{\sigma}^2_t) + \frac{\hat{\sigma}^2_{t,prox}}{\tilde{\sigma}^2_t}\right), & \text{(QLIKE)} \end{aligned} \tag{37}$$

where the robustness of these functions is shown by Patton (2011). Robustness of a loss function means that the rank of a volatility model in equation (27) is equivalent when using a proxy $\hat{\sigma}^2_{t,prox}$ and when using the 'true' unobserved volatility σ^2_t. In figure 1 in the article by Patton (2011), the shapes of these different loss functions $L(\cdot, \cdot)$ are plotted. It can be seen that the MSE is a symmetric loss function that penalizes positive and negative errors similarly, whereas the QLIKE loss function punishes positive outliers less than negative outliers. This characteristic is particularly useful in volatility model evaluation, as accounting for higher volatility can be seen as a *margin of safety*, while accounting for lower volatility than what is actually observed can lead to excessive risk taking.

Subsequently, Hansen and Lunde (2006) and Patton (2011) discuss the optimal choice of a volatility proxy $\hat{\sigma}_{t,prox}$. The most common choice for $\hat{\sigma}^2_{t,prox}$ are the squared daily returns, under the assumption that the conditional mean of the returns is zero. Since the conditional mean of r_t is argued to follow an ARMA process described by equation (16), this does not hold in this paper. Throughout their papers, it is argued that the *realised volatility* (the sum of squared intraday returns) is the best proxy for volatility, however this requires intraday data which is not available for this paper[7]. Therefore, the squared residuals ε_t of the benchmark model are used as a proxy for conditional volatility, since it is easily checked that $E[\varepsilon^2_t \,|\mathcal{F}_{t-1}] = \sigma^2_t$, where once again $\mathcal{F}_{t-1}$ denotes the information set at time t.

Hence, using ε^2_t from equation (16) as proxy for the conditional volatility, the in-sample fit of the introduced models can be compared. Calculating the values of the MSE and QLIKE loss functions yields the MSE and QLIKE error for a conditional volatility process defined by $\tilde{\sigma}^2_t$. This allows to rank the processes $\tilde{\sigma}^2_t$ defined by equation (27) under the constant and varying-parameter specification, to make a qualitative assessment about the in-sample fit of these models.

[7]The WRDS license of Tilburg University does not provide access to the TAQ database.

5 Empirical setup

To test the effect of Twitter-based exogenous variables on GARCH-type volatility models proposed in section 4.1, the methodologies are applied to real-world data. This chapter discusses the data that is used for this research, provides descriptive statistics, and provides relevant time series plots. Furthermore, the appropriate data-driven choice for the sentiment models from section 3.2 is determined and the exogenous GARCH processes are calculated. The Python code of the empirical study can be accessed via GitHub[8].

5.1 Data

This section discusses how the companies included in this research are selected. Subsequently, a summary of the evaluation metrics of the introduced sentiment models is provided, and the most suitable method for this research is determined. Lastly, this section discusses relevant variables that are suited to enter the equation from equation (34), which are used to account for possible shifts in the volatility process.

Before the Twitter data is investigated, the granularity of the data and time period of this research must be established. In the literature, there are sometimes confronting views about the effect of Twitter-based exogenous variables on stock prices. Barberis, Schleifer, and Vishny (1998) argue that the incorporation of sentiment in stock prices only happens gradually, whereas Audrino et al. (2020) conclude that sentiment and attention variables have predictive power for up to two-day ahead predictions. As the models from section 4.1 allow for the inclusion of longer lags of the exogenous variables, daily data is used. Using daily data, both long-term and short-term effects of Twitter on the volatility of stock price return can be discovered. Following Lamoureux and Lastrapes (1990), it is expected that using high-frequency data increases the total degree of persistence.

To account for both possible time-varying effects as proposed in section 4.4 and the gradual adoption of Twitter as a news and sharing platform, a horizon of 10.5 years is proposed, starting from 01-01-2011 until 31-08-2021. This allows for discovering possible asymmetric and time-varying effects as multiple recessions occurred during this time period. After 2010 social media platforms and the online culture became increasingly adopted throughout our culture.

5.1.1 Company selection

This research investigates the effect of Twitter-based sentiment metrics on stock price movements of U.S. companies as both the U.S. stock markets (NASDAQ and NYSE) are the most high profile and as the U.S. population active on Twitter is the largest. To accurately model the sentiment and get a weighted average sentiment score, a sufficient amount of Twitter coverage per company is crucial. Subsequently, the U.S. companies selected in this research must generate a sufficient amount of tweets.

Accordingly, among the 50 largest companies in the S&P 500 based on Market Capitalization[9] a Twitter search is conducted. For all these companies, the relevant tweets obeying the search query from section 3.1, send between the 1st and 31st of May 2021 are selected, and companies with limited coverage (less than 25 tweets per day on average) are filtered out[10]. This is done in order to not include companies in this research with limited Twitter coverage since for these companies the effects of Twitter on stock price return volatility are possibly less evident. Note that the negation dictionary is only constructed after the initial company selection, hence, this initial search does not make use of a negation dictionary. Although this

[8] Full code (via GitHub)

[9] At 01-09-2021

[10] Note that the Twitter API only allows retrieval for 10 million tweets per month, which makes retrieving tweets for all 50 companies for an extended period impossible.

is likely to increase the number of spam tweets in the initial search, it provides valuable insights into the number of tweets sent per day.

After filtering out companies that do not generate a sufficient amount of Twitter attention in May 2021, 14 companies are left and are included in this research. Based on the tweets found in the initial Twitter search, a company specific negation dictionary is constructed, in line with the explanation from section 3.1. Although most tweets retrieved were relevant for this research, in the case of Amazon, Apple, McDonalds, Google, Facebook, Microsoft, and NIKE, a negation dictionary is necessary to filter out unwanted tweets. More specifically, some of the retrieved tweets refer to products these companies sell (*e.g.* 'green Nike Air Max'), and do not express emotions or opinions that might influence stock price volatility. Furthermore, there are multiple product links referring to products on Amazon or podcast/shows on Apple TV, therefore, common terms in these tweets are included in the negation dictionary. A quick manual inspection from the data for Facebook and Google learns that a substantial amount of tweets in the dataset is not related to the business practices of Facebook Inc and Alphabet Inc., respectively. Even after using the negation dictionary, the data mostly consists of spam tweets, hence, these companies are omitted from this research.

Table 1 show the companies that are included in this research. From the table, the selected companies are very diverse in their business model and their maturity. This might lead to differences in the perceived effects of Twitter on stock price companies across industries, which is interesting to explore. Additionally, table 1 reports the number of tweets for each company included in this research.

Table 1 – Selected U.S. companies

Company name	Number of *tweets*	Company name	Number of *tweets*
Apple (AAPL)	2,049,809	Microsoft (MSFT)	1,059,333
Amazon (AMZN)	1,024,467	Netflix (NFLX)	285,626
Chevron (CVX)	80,494	NIKE (NKE)	212,297
Coca-Cola (KO)	218,282	salesforce (CRM)	153,085
Exxon Mobil (XOM)	49,659	Tesla (TSLA)	399,564
McDonalds (MCD)	398,542	Walmart (WMT)	683,628

Companies from the S&P500 included in this research. Reported are the number of tweets between 01-01-2011 until 08-31-2021 obeying the query with the *ESG-dictionary* and manually constructed *negation dictionary*. Ticker symbol of the NYSE or NASDAQ stock exchange for each company is provided in parentheses.

Besides the different levels of maturity and different types of business, other factors can be identified that cause different effects of Twitter-based sentiment metrics on stock price volatility. Audrino et al. (2020) observe the largest increase in predictive accuracy from including sentiment and attention data for companies with large market capitalization and/or low percentage of institutional investors. Specifically, stock volatility from stocks that are largely held by retail investors are more affected by public sentiment. This supports the intuition that retail investors tend to be more affected by sentiment than institutional investors. For the companies in this research, the market capitalization and percentage of shares held by institutional investors are reported in table 13 in appendix A.1.2. Following the conclusions by Audrino et al. (2020) and Rakowski et al. (2021), it is expected that for Apple, Chevron, Exxon Mobil, McDonalds, and Tesla, Twitter-based variables can significantly explain the conditional volatility of the stock price returns. Whereas it is expected that for Netflix, NIKE, salesforce, and Walmart, significant effects of Twitter are less frequently observed.

5.1.2 Sentiment classification

In this subsection, the sentiment classification methods from section 3.2 are tested on a manually annotated subset of the gathered Twitter data. In total, a panel of eight students classified a subsample of 2200 tweets from the gathered tweets shown in table 1. These manually determined sentiment polarity scores are regarded as *ground-truth*. The distribution of instance in class j is approximately equal, which indicates that for every class j, approximately one third of the tweets are annotated with label j. The objective of this section is to determine the method that achieves the best overall rank with regard to the proposed evaluation metrics by equation (10) and equation (11). After the classification by the regular VADER lexicon, the falsely classified items from the test-set are manually inspected, and domain specific items are added to the dictionary to improve the performance of the VADER sentiment classification method. As aforementioned, overfitting must be avoided, therefore the items that are added to the lexicon are mostly words from the ESG dictionary from table 12 in appendix A.1.1, assigned a polarity score. As the scores from the VADER sentiment lexicon by Hutto and Gilbert (2015) are extensively reviewed by a panel of reviewers, determining the score for each introduced items individually is a tedious task, while determining the score via a data-driven approach introduces the risk of overfitting to the small subsample. Therefore, the score is assumed to be equal for all introduced words, and only the sign determines the polarity and influence the sentiment classification. The numerical score attached to all introduced items is the score that maximizes the accuracy and macro F_1 score, which is 1.5.

The adjusted VADER lexicon allows for correct classification of some very domain specific issues. For instance, a very objectively worded tweet discussing the large role of company X in climate change related issues would normally be classified as neutral, whereas these tweets are argued to generate a negative sentiment of the public towards this company. The following tweet about NIKE is now classified correctly despite its objective tone of voice

> *Nike & IKEA invest to scale waterless textile dyeing system to make #water-intensive industry more sustainable http://t.co/7E95UGGqUn* - @GlobalSherpa, 15-05-2015.

Recall that before using the VADER lexicon to classify the tweet, the only preprocessing step was to remove hashtags, mentions, and links, which leads to the following tweet

> *Nike & IKEA invest to scale waterless textile dyeing system to make industry more sustainable.*

This tweet very objectively informs other Twitter users of the sustainable move that NIKE has undertaken. However, as sustainable is a very domain specific word, it does not have a score in the VADER lexicon, which classifies this tweet as neutral. By extending the VADER lexicon and assigning *sustainable* a positive polarity (with score +1.5), this tweet is correctly classified as positive by the adjusted VADER lexicon.

Table 2 – Performance metrics of sentiment polarity classifiers.

	Naive Bayes	VADER	Adjusted VADER
Macro F_1	0.377	0.462	**0.474**
Accuracy	0.387	0.47	**0.489**

Macro F_1 and accuracy score of the NB, VADER and Adjusted VADER multi-class sentiment classification methods. Goal of the classification algorithm is to correctly classify instances as positive, neutral or negative (three-class classification). Metrics are calculated on a manually classified *ground-truth* subsample of the Twitter datasets. Numbers in bold denote the best score of a metric among the classification methods.

Subsequently, the Macro F_1 score and accuracy for the proposed models on the manually annotated test set is calculated. The results are provided in table 2. As aforementioned, that the Naive Bayes Classifier was trained on a publicly available labeled training set, where all domain irrelevant words[11] were removed to increase the performance of this model. From table 2, it can be seen that across all sentiment classification method, the Adjusted VADER method, tuned according to the very specific domain of the retrieved tweets, performs best across the two complementary metrics. However, this method is shown to provide little benefit over the use of the regular VADER lexicon. Lastly, it can be noted that across all sentiment classification methods the score for both evaluation metrics is very low, specifically in the case for NB, where the accuracy is only slightly higher than would be expected by pure chance. This can be attributed to the fact that despite although the training data consists of tweets, they are widely different from the tweets in the test set. The low accuracy of the multi-class sentiment classification is in line with the observation by Hartmann et al. (2019), who identifies low accuracy in multi-class classification assignments. As the Adjusted VADER achieves the highest score for both metrics, this method is used in this research to determine the polarity score of the tweets.

However, it must be noted that a substantial fraction of the tweets is incorrectly classified, and 10%-20% of the tweets in the test set is actually assigned the opposite polarity label. This might indicate that the calculated Twitter-based sentiment score is not an actual representation of the sentiment among Twitter users, and it fails to meaningfully explain volatility of stock price returns. When the sentiment score does not successfully explain volatility, this could be attributed to both the failure of the correct classification of sentiment, as well as there is simply no effect of Twitter-based sentiment on conditional volatility.

5.1.3 Volatility drivers

As discussed in section 4.4, this research additionally explores whether the effect of Twitter-based variables is dependent on the state of the economy. By taking into account macroeconomic or market conditions, the robustness of the parameters are investigated and time-varying effects can be found. To take into account the possible time-varying effects on stock price volatility, the parameters of the GARCH alternatives are allowed to shift over time based on a set of time series $\boldsymbol{\zeta}_t$. This subsection discusses in greater extent possible choices to be included in this set which must be available on a daily basis.

Mittnik, Robinzonov, and Spindler (2015) aim to find key volatility drivers by exploiting a gradient boosting method to find both financial and macroeconomic factors that seem to explain volatility. Their methodology of including these risk drivers proves to increase forecasts. By including 40 different financial and macroeconomic factors, classified in different categories their aim is to find the volatility drivers that lead to the most increase in predictive ability. Their methodology provides more useful than the usual GARCH framework for this particular task, however, the derived volatility drivers can be used in a GARCH framework likewise.

In their research, Mittnik et al. (2015) find that the CBOE Volatility Index is an important predictor for realized volatility. This index is across all horizons preferred by their selection algorithm. The CBOE Volatility Index is created by the Chicago Board Options Exchange (CBOE) to measure the market's expectation of future volatility (Kuepper, 2021) and is often simply called 'the VIX'. It is derived from the prices of options of the S&P500 index options, with close maturity dates (between 23 and 37 days ahead). Therefore, the VIX generates a 30-day forecast of market volatility, which is often considered a way to observe market sentiment, and in particular the degree of fear among investors (Kuepper, 2021). Specifically, high values of the VIX correspond to fearful investors, who are afraid of a market crash,

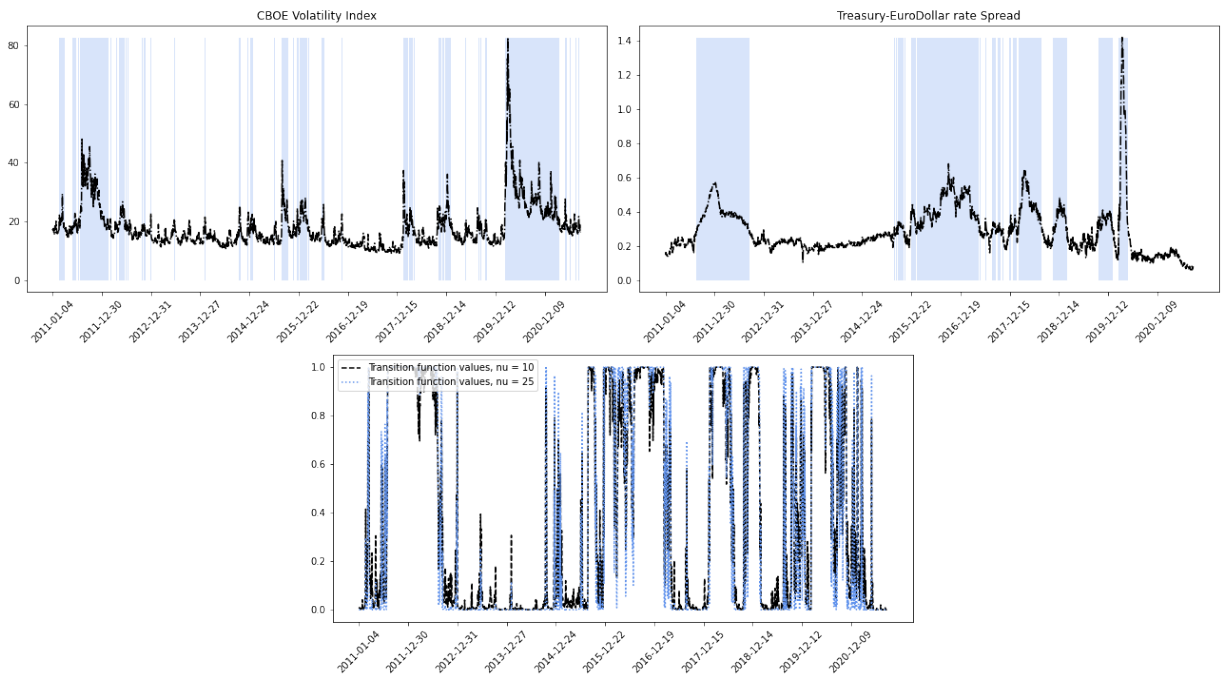
[11] These words are associated with the source of the training data. The words that are removed are 'aapl', 'apple', 'flight', 'airline', 'airtravel', 'airpassenger', 'delayed', 'gate', and 'terminal'.

expecting worsening market conditions, and low values correspond to *bullish* investor belief. The VIX is a real-time index and is therefore available on a daily basis.

Furthermore, Mittnik et al. (2015) find that from the macroeconomic variables investigated that are available on a daily basis, only the Treasury-EuroDollar rate (TED) spread is not excluded as volatility driver by their algorithm. Subsequently, this research also includes the TED spread as control factor. The TED spread is an indicator of credit risk, and is the difference in basis points between the three-month LIBOR (London Interbank Offered Rate) and three-month U.S. Treasury bill rate. More specifically, the TED spread is the difference between the rate of short term interbank loans and the interest rate on short term US government debt (J. Chen, 2020). In a weakened economy banks perceive loans to corporate clients more risky, and in turn charge higher interest rates on these loans. When the LIBOR rate (rate on corporate loans) deviates more from the risk-free U.S. government rate on T-bills, the TED rate rises. Therefore, the TED spread signals a weakening economic strength.

Mittnik et al. (2015) state that the VIX and the TED spread are able to signal both increasing and decreasing volatility while other variables were only able to signal increasing volatility. To account for possible regime shifts, define $\boldsymbol{\zeta}_t = (VIX_t, TED_t)$ as the variables that drive the smooth transition operator discussed in section 4.4. Figure 1 provides a time series plots of these metrics.

Figure 1 – Volatility drivers and regimes



Time series plots of the CBOE Volatility Index (VIX) (left) and Treasury-EuroDollar (TED) rate (right) spread between 01-01-2011 and 31-08-2021 (upper graphs). Shaded areas correspond to periods where the VIX/TEDRATE exceeds its sample mean, which might indicate worsening conditions in the market/economy. Lower graph displays values of the logistic transition function for $\nu = 10, 25$ based on the standardised[12]values of the VIX and TEDRATE.

Since these variables can be regarded as a proxy for investor sentiment in the market and weakening economic strength, respectively, higher values for both variables correspond to worsening economic conditions, whereas low values correspond to economic prosperity. Therefore, no adjustment to these variables have

[12]Let X_t be a time series process for $t = \{1, \ldots, T\}$, then the standardized value of X_t is given by $\frac{X_t - \bar{X}}{std(X)}$, where $\bar{X}, std(X)$ denotes the sample mean and sample standard deviation of X, respectively.

to be made before they can enter the function from equation (34). Note that this function is one under worsening economic conditions and increased fear in the market, and zero when investors and banks are confident about the economy/market.

From figure 1 can be seen that although some periods of worsening macroeconomic or market conditions overlap, there are some difference where the VIX and the TED rate exceed their respective sample means. Both time series show increasing values around the COVID-19 outbreak, and around the sovereign debt crises in Europe (Greece). Furthermore, it is remarkable to notice that in the case for the TED rate the spike relating to the COVID-19 outbreak was relatively short, whereas increased fear in the market remained until the end of 2020. In other periods, it can be seen that the TED rate seems to be more gradually increasing, whereas the VIX consists of more spikes, for instance for the period 2017/2018. In line with Hsu et al. (2021) and Smales (2014), it is expected that during economic distress (*i.e.* when $G(\cdot) = 1$), Twitter-based variables have a more significant effect on the conditional volatility of stock price returns. Furthermore, based on Smales (2014), it is expected that this effect is asymmetric, such that the effect of negative Twitter-based sentiment has more effect on the conditional volatility than negative Twitter-based sentiment.

The lower graph in figure 1 shows time series plots of equation (34) for different values of ν, which indicates spikes in the standardized VIX and TED rate and thus signals increased fear in the market. It can be seen that for a lower value of ν, the smooth transition function is much more sensitive to small deviations from the mean of the variables in $\boldsymbol{\zeta}_t$. As discussed in section 4.4, ν is chosen as the value that generally maximizes the Quasi Log Likelihood of Model II across different companies for $\nu \in [10, 15, 25, 50, 100]$. While there is no consensus across companies for the optimal neighborhood of ν based on the range of possible values, the reported maximized likelihoods across the possible values are declining or increasing with ν for half of the companies. For the other companies, the optimal smooth transition operator is in the neighborhood of the center of the grid. However, the differences in maximum Quasi Log Likelihood for different values of ν are almost negligible. To ensure the same behaviour of the varying parameters based on the function $G(\cdot)$, $\nu = 25$ for all companies.

5.2 Descriptive statistics

In this section, the descriptive statistics of the company stock data, Twitter-based sentiment data and the control variable data is given. Additionally, the parameter estimates of the exogenous GARCH process, described by equation (21) and equation (22) are given, which serve as exogenous variables in the stock price volatility model from equation (23).

5.2.1 Market data

For the period 2011-01-01 until 2021-08-31 daily stock prices are retrieved via Yahoo! Finance. Stock price returns are characterized by $r_t = 100 \cdot \log\left(\frac{S_t}{S_{t-1}}\right)$, where S_t denotes the closing price at time t. Before the market based exogenous variables can enter the model, they must be standardized, otherwise their highs and lows are not comparable. Define the standardized value of a time series variable X_t by $X_{stand,t} = (X_t - \bar{X})/s$, where $s = \sqrt{\frac{\sum_{i=1}^{T}(X_t - \bar{X})^2}{T-1}}$ and $\bar{X} = \frac{1}{n}\sum_{t=1}^{T} X_t$.

Table 3 provides the descriptive statistics of the return time series of the companies in the dataset, as well as the macroeconomic and market indicator time series, the TED rate and the VIX. From the table it can be seen that the median, and the 75% and 25% quantiles are comparable across the companies. The standard deviation, sample minimum and sample maximum are very different across the selected companies. When looking at the range of values and the standard deviation of NFLX and TSLA it can be seen that the sample minimum and maximum (in absolute values) and sample standard deviation is much higher compared to

Table 3 – Descriptive statistics of the returns of U.S. companies and of the selected volatility drivers.

	Stock returns of selected companies												***Volatility drivers***	
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT	VIX	TEDRATE
Mean	0.10	0.11	0.00	0.02	-0.01	0.04	0.09	0.12	0.08	0.08	0.18	0.04	0.00	0.00
Std.	1.79	1.96	1.76	1.11	1.57	1.22	1.60	3.15	1.65	2.18	3.47	1.21	1.00	1.00
Min	-13.77	-13.53	-25.01	-10.17	-13.04	-17.29	-15.95	-42.92	-12.41	-17.30	-23.65	-10.74	-1.17	-1.57
25%	-0.73	-0.82	-0.76	-0.47	-0.72	-0.48	-0.66	-1.22	-0.71	-0.92	-1.48	-0.52	-0.62	-0.61
50%	0.09	0.12	0.04	0.03	-0.01	0.07	0.07	0.03	0.07	0.10	0.12	0.05	-0.27	-0.26
75%	1.03	1.12	0.78	0.57	0.72	0.59	0.88	1.50	0.89	1.16	1.88	0.61	0.26	0.56
Max	11.32	14.62	20.49	6.28	11.94	16.66	13.29	35.22	14.44	23.15	21.83	11.07	8.88	7.80

Descriptive statistics of the selected companies for this research in the period 01-01-2011 until 31-08-2021 ($T = 2683$). Companies are described by their ticker symbol. On the right, the descriptive statistics of the standardized values of the volatility drivers from section 5.1.3 are given as well.

other companies. This may indicate higher or more frequent outliers in the volatility process. Lastly, by observing the selected volatility drivers, it can be seen that the magnitude of the sample maximum is much higher than sample minimum for both variables, which indicates that both of these time series contain high spikes, which is supported by the time series plots in figure 1.

5.2.2 Exogenous variables

For each company included in this research, the exogenous variables that enter the model are described by the daily sentiment score from equation (15), the daily number of tweets, and the number of interactions. As these variables enter the conditional volatility models, these process must satisfy the stationarity assumption, for which the Augmented Dickey Fuller test is used as explained by section 4.2. If the null hypothesis can not be rejected, the log difference of the time series is used in the analysis, in this case the corresponding differences time series will be displayed as Δx_t. For all Twitter-based exogenous variables the null hypothesis that the time series process had a unit root was rejected at the 5% significance level. Therefore, these variables can directly enter the models proposed in section 4.1.

Recall that the ARMA-apARCH-apXGARCH model from equation (23) assumed that the exogenous variables themselves also followed an ARMA-GARCH process. The volatility of this process, $\sigma_t^{(x)}$ is subsequently used to explain the conditional volatility of stock price returns. Under the assumptions from section 4.2, the parameter estimates that minimize equation (29) form a time series of exogenous conditional variances, since these variables are non-negative and stationary by definition, these series can directly enter the conditional volatility equation. Minimizing[13] the Quasi Log Likelihood function results in the parameter estimates that create this series.

Table 4, table 5, and table 6 display the descriptive statistics and the optimal parameters of the time series process of the daily Twitter sentiment, daily number of tweets and daily number of interactions, respectively. Since the Twitter-based variables are standardized, the mean of the exogenous series is almost equal to zero and the standard deviation equal to one. By inspecting the sample minimum and maximum, it can be seen that in the case for the Twitter derived sentiment, the sample minimum and maximum are symmetrically distributed around the sample mean. However, this does not hold for the number of tweets and the number of interactions, where the sample mean is for most companies close to zero and the sample maximum is in most cases much larger than 10 times the standard deviation, indicating a

[13] minimization is done using the *Sequential Least Squares Programming* (SLSQP) minimization in the *SciPy* library in Python.

skewed distribution and lots of spikes in the time series. If these positive outliers correlate with spikes in the volatility, the attention variable could convey a lot of information and serve as an relevant variable to explain the spikes in the volatility process.

Table 4 – ARMA(1,1)-GARCH(1,1) parameter estimates and descriptive statistics of Twitter sentiment time series from 01-01-2011 until 31-08-2021.

	$\hat{\varphi}_x$ ***of ARMA(1,1)-GARCH(1,1) process of Twitter sentiment***											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
μ_x	0.0	-0.0*	0.004***	-0.0	0.086***	-0.019***	0.002***	-0.0***	0.001***	-0.0	-0.0***	0.001***
	(4.3e-04)	(2.7e-05)	(4.1e-05)	(1.3e-04)	(4.1e-03)	(1.3e-05)	(8.7e-05)	(8.3e-05)	(1.5e-04)	(1.9e-03)	(4.4e-05)	(9.4e-08)
γ_x	0.6***	0.549***	0.776***	0.822***	-0.948***	0.59***	0.679***	0.984***	0.906***	0.989***	0.967***	0.181
	(2.8e-02)	(3.1e-02)	(1.7e-01)	(4.9e-02)	(6.1e-02)	(9.7e-02)	(2.7e-02)	(6.3e-04)	(6.9e-03)	(4.9e-04)	(2.6e-03)	(4.0e+00)
δ_x	-0.401***	-0.361***	-0.638***	-0.689***	0.96***	-0.413***	-0.457***	-0.921***	-0.804***	-0.916***	-0.883***	-0.031
	(5.8e-06)	(4.2e-05)	(9.2e-03)	(2.1e-02)	(2.0e-02)	(6.3e-05)	(2.1e-05)	(1.7e-04)	(1.2e-04)	(6.8e-04)	(3.6e-02)	(3.9e+00)
ω_x	0.281***	0.691***	0.024***	0.164***	0.002***	0.022***	0.762***	0.373***	0.475***	0.257***	0.201***	0.075***
	(9.3e-12)	(4.5e-11)	(2.5e-10)	(3.3e-11)	(7.8e-09)	(4.2e-10)	(1.7e-09)	(5.5e-07)	(2.6e-08)	(6.8e-09)	(9.2e-09)	(2.4e-13)
β_x	0.647***	0.195***	0.958***	0.784***	0.983***	0.941***	0.113***	0.512***	0.427***	0.558***	0.665***	0.848***
	(2.9e-04)	(2.8e-02)	(1.4e-02)	(2.4e-02)	(1.9e-02)	(8.6e-04)	(3.6e-02)	(1.5e-04)	(2.0e-04)	(1.1e-04)	(2.7e-02)	(4.4e-04)
α_x	0.056***	0.082***	0.017***	0.044***	0.015***	0.037***	0.058***	0.08***	0.07***	0.132***	0.12***	0.078***
	(1.3e-03)	(2.1e-02)	(7.4e-04)	(4.7e-03)	(5.5e-04)	(2.6e-03)	(3.9e-03)	(1.3e-03)	(4.7e-03)	(1.4e-03)	(1.8e-03)	(1.8e-02)
	Descriptive statistics											
Mean	1.66e-17	1.74e-16	2.73e-16	6.34e-16	-1.65e-16	-1.19e-16	-7.39e-17	-5.62e-16	4.90e-17	-5.40e-15	5.20e-16	3.01e-17
Std.	1.00e+0	1.00e+0	9.95e-1	1.00e+0	9.02e-1	1.00e+0	1.00e+0	1.00e+0	1.00e+0	9.97e-1	9.98e-1	1.00e+0
Min	-5.05e+0	-6.46e+0	-4.13e+0	-4.13e+0	-4.71e+0	-4.59e+0	-5.33e+0	-4.54e+0	-5.53e+0	-3.79e+0	-4.87e+0	-6.18e+0
Max	5.17e+0	4.31e+0	4.76e+0	4.00e+0	6.61e+0	5.22e+0	3.56e+0	4.39e+0	4.14e+0	3.09e+0	3.28e+0	4.77e+0
$\mathcal{L}_T$	-7444.05	-7467.07	-7432.23	-7457.75	-6382.80	-7362.09	-7369.99	-7339.12	-7437.85	-7038.38	-7297.05	-7445.66

Parameter estimates of Twitter derived daily sentiment score time series. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1%, respectively. In the lower rows the descriptive statistics of the standardized daily sentiment series for each company is provided, which directly enters the conditional volatility process. Lastly, the optimized Quasi Log Likelihood is given.

Moreover, the tables provide the parameter estimates, standard errors and the reported significance of the proposed ARMA(1,1)-GARCH(1,1) structure. The t-test from equation (35) is employed to calculate the significance of the parameters at the 10%, 5% and 1% level. From the tables it can be inferred that for almost all companies the parameters describing the process of the Twitter-based Twitter-based exogenous variables are significant. In most cases (except for the daily number of interactions for XOM and NFLX), the null that no ARCH structure is present can be rejected, as $\alpha_x \neq 0$ at the 1% significance level. Across all companies and exogenous variables, the majority of the parameters is tested to be significant. The ARMA-GARCH process for each variable seems to be highly persistent across all companies. To clarify, in almost all cases $\beta_x + \alpha_x$ is close to one, indicating long term effects of unexpected values of the exogenous variables.

The volatility of the exogenous variables are visualized in figure 3, figure 4, and figure 5 in appendix A.2.1. These plots provide an intuition into whether the disagreement variables $\sigma_t^{(x_k)}$ can successfully explain the conditional volatility of stock price returns. From the plots, it is seen that the volatility of the Twitter-based sentiment is generally represented by a flat line, albeit minor fluctuations are perceivable. Especially in the case for Apple, Amazon, Microsoft, Netflix, and NIKE it is expected that the volatility of the Twitter-based sentiment fails to explain spikes in the stock price return volatility. For Exxon Mobil, McDonalds, salesforce, and Tesla, the volatility of the Twitter-based sentiment shows more variation, and thus the value of $\sigma_t^{(x)}$ might help to explain conditional volatility of stock price returns. Now, consider the volatility of the attention variables. The plots in appendix A.2.1 show that only for Apple, Microsoft, NIKE, and Tesla, the volatility of the standardised number of tweets does not account for high spikes. The plots of the other companies look more promising, although the spikes in the volatility of the stock price returns must align with the spikes in the volatility of the number of tweets for these variables to have any influence. In the case for Apple, Exxon Mobil, Microsoft, and Netflix, it is evident that the volatility of the standardised number of interactions likely will not successfully explain the volatility of stock price returns, since $\sigma_t^{(x)}$ represents a flat line for these companies. For the other companies, the volatility of the standardised number of interactions can explain increased levels of stock price return volatility when the spikes correlate.

Table 5 – ARMA(1,1)-GARCH(1,1) parameter estimates and descriptive statistics of daily number of tweets from 01-01-2011 until 31-08-2021.

	$\hat{\varphi}_x$ ***of ARMA(1,1)-GARCH(1,1) process of daily # of tweets***											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
μ_x	-0.222***	-0.463***	-0.039**	-0.389**	-0.035***	-0.013	0.002***	0.006***	0.001***	0.004***	-0.223***	-0.009***
	(2.7e-05)	(2.3e-02)	(1.8e-02)	(2.4e-01)	(4.5e-04)	(2.1e-02)	(2.2e-06)	(9.3e-04)	(1.0e-05)	(9.0e-06)	(9.2e-07)	(3.9e-05)
γ_x	0.238**	-0.488***	0.723***	-0.927**	0.263*	0.496**	0.845***	0.981***	0.784***	0.921***	0.527***	0.867***
	(1.4e-01)	(3.5e-02)	(8.6e-02)	(5.3e-01)	(1.9e-01)	(2.7e-01)	(1.7e-02)	(5.2e-02)	(6.6e-02)	(2.6e-02)	(1.9e-01)	(6.0e-03)
δ_x	0.076	0.818***	-0.938***	0.293*	0.44***	-0.276***	-0.641***	-0.798***	-0.574***	-0.598***	-0.025	-0.728***
	(1.2e-01)	(5.7e-04)	(4.2e-02)	(2.2e-01)	(2.3e-05)	(4.6e-02)	(2.7e-04)	(1.8e-01)	(1.1e-04)	(2.4e-05)	(1.2e-01)	(2.2e-04)
ω_x	0.006***	0.195***	0.09***	0.15***	0.08***	0.592***	0.249***	0.039***	0.579***	0.303***	0.001***	0.145***
	(6.1e-11)	(7.1e-06)	(1.1e-04)	(4.1e-06)	(1.4e-08)	(3.4e-12)	(5.3e-09)	(8.8e-07)	(2.5e-10)	(3.2e-09)	(6.3e-13)	(2.1e-07)
β_x	0.947***	0.525***	0.016	0.001	0.577***	0.0	0.643***	0.793***	0.293***	0.32***	0.982***	0.597***
	(5.5e-03)	(2.7e-04)	(6.7e-01)	(5.7e-02)	(1.0e-03)	(1.1e-06)	(7.1e-04)	(1.8e-01)	(7.1e-02)	(2.8e-02)	(6.7e-03)	(5.5e-04)
α_x	0.053***	0.475***	0.984***	0.999***	0.423***	0.447***	0.078***	0.207***	0.05***	0.31***	0.018***	0.189***
	(2.4e-03)	(4.2e-02)	(2.0e-02)	(4.1e-02)	(3.3e-02)	(7.1e-02)	(4.7e-03)	(4.4e-03)	(1.0e-02)	(2.9e-02)	(1.3e-03)	(4.4e-03)
	Descriptive statistics											
Mean	-2.65e-16	-4.77e-16	1.01e-17	-1.22e-17	7.89e-17	-2.22e-17	2.04e-16	-1.24e-15	1.22e-16	4.15e-16	9.95e-16	6.55e-17
Std.	1.00e+0	1.00e+0	9.95e-1	1.00e+0	9.02e-1	1.00e+0	1.00e+0	1.00e+0	1.00e+0	9.97e-1	9.98e-1	1.00e+0
Min	-9.58e-1	-8.94e-1	-2.87e-1	-3.04e-1	-2.69e-1	-6.31e-1	-1.03e+0	-6.89e-1	-7.16e-1	-8.93e-1	-6.14e-1	-7.37e-1
Max	1.39e+1	1.21e+1	3.38e+1	2.00e+1	2.72e+1	2.19e+1	1.20e+1	2.56e+1	1.74e+1	2.06e+1	1.86e+1	2.21e+1
$\mathcal{L}_T$	-6620.15	-7175.38	-6089.41	-4160.47	-4444.82	-6667.26	-7145.90	-5658.14	-7193.36	-6161.08	-5581.48	-5875.98

Parameter estimates of the daily number of tweets. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1%, respectively. In the lower rows the descriptive statistics of the standardized daily number of tweets for each company is provided, which directly enters the conditional volatility process. Lastly, the optimized Quasi Log Likelihood is given.

The time series of the exogenous variables, as well as the residuals $\epsilon_t^{(x)}$ from the autoregressive moving-average structure imposed on the exogenous variables, are evaluated for each company. Via partial-autocorrelation plots, it is checked whether the imposed ARMA-GARCH structure is sensible for the exogenous variables for all companies. Across all companies, these plots did not show significant evidence that higher order lags would hold any predictive power. However, it must be noted that in contrast to the stock return models, the proper lag order specification of the exogenous processes was not evaluated by an evaluation criterion (*e.g.* BIC), which might lead to patterns that are not discovered by the proposed ARMA(1,1)-GARCH(1,1) specification, or cases where the proposed model is too restrictive.

Table 6 – ARMA(1,1)-GARCH(1,1) parameter estimates and descriptive statistics of daily number of interactions from 01-01-2011 until 31-08-2021.

	$\hat{\varphi}_x$ ***of ARMA(1,1)-GARCH(1,1) process of daily # of interactions***											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
μ_x	-0.261***	-0.162***	-0.031***	-0.185	-0.003	-0.018	-0.287***	-0.064*	-0.225***	-0.256***	-0.021	-0.006*
	(4.0e-02)	(1.6e-07)	(7.8e-04)	(1.8e-01)	(4.4e-03)	(1.8e-02)	(1.9e-04)	(4.4e-02)	(1.1e-02)	(4.0e-05)	(8.3e-02)	(5.0e-03)
γ_x	-0.098***	0.098***	0.743***	-0.698***	0.914***	0.797***	0.248*	-0.134	-0.927***	0.213	0.918***	0.946***
	(2.3e-02)	(2.2e-02)	(2.8e-01)	(1.2e-02)	(2.5e-01)	(8.0e-02)	(1.6e-01)	(8.9e-01)	(2.2e-02)	(2.6e-01)	(9.5e-02)	(9.8e-04)
δ_x	0.58***	0.476***	0.853***	0.988***	-0.921***	-0.615***	0.09	0.142	1.0***	0.104	-0.976***	-0.99***
	(2.1e-04)	(1.7e-04)	(4.2e-03)	(9.0e-02)	(4.8e-04)	(9.7e-05)	(2.1e-01)	(8.2e-01)	(6.3e-02)	(2.6e-01)	(1.0e-01)	(9.0e-02)
ω_x	0.0***	0.0***	0.0***	0.0***	0.004***	0.871***	0.0***	0.001***	0.0***	0.0***	0.0	0.005***
	(2.6e-09)	(2.2e-13)	(1.4e-09)	(2.7e-07)	(1.1e-16)	(1.5e-07)	(4.8e-13)	(4.2e-21)	(5.5e-08)	(1.6e-11)	(1.2e-05)	(3.7e-06)
β_x	0.993***	0.966***	0.797***	0.894***	0.996***	0.007***	0.992***	1.0***	0.959***	0.991***	0.969***	0.003
	(3.2e-04)	(5.6e-04)	(4.7e-02)	(1.1e-02)	(4.7e-03)	(1.8e-03)	(8.8e-06)	(1.0e-03)	(9.3e-02)	(3.4e-03)	(1.6e-03)	(9.0e-03)
α_x	0.007***	0.034***	0.203***	0.106***	0.0	0.993***	0.008***	0.0	0.041***	0.009***	0.031***	0.997***
	(1.1e-04)	(2.2e-07)	(1.4e-04)	(5.0e-05)	(2.0e-06)	(2.0e-01)	(5.1e-05)	(1.0e-05)	(3.4e-07)	(8.9e-04)	(1.1e-06)	(6.3e-04)
	Descriptive statistics											
Mean	-1.31e-16	6.49e-16	3.52e-16	2.01e-16	2.05e-17	2.25e-16	9.39e-16	5.50e-17	9.01e-17	-6.59e-17	-2.78e-16	-1.32e-16
Std.	1.00e+0	1.00e+0	9.95e-1	1.00e+0	9.02e-1	1.00e+0	1.00e+0	1.00e+0	1.00e+0	9.97e-1	9.98e-1	1.00e+0
Min	-2.75e-1	-1.81e-1	-1.35e-1	-1.21e-1	-9.57e-2	-1.54e-1	-4.05e-1	-6.81e-2	-1.19e-1	-3.56e-1	-2.61e-1	-1.48e-1
Max	3.28e+1	3.41e+1	3.57e+1	3.93e+1	4.25e+1	2.39e+1	2.51e+1	4.94e+1	4.24e+1	1.60e+1	2.82e+1	2.98e+1
$\mathcal{L}_T$	-4452.83	5875.82	246.39	-426.85	-6694.78	-7511.90	-2163.265	-6023.16	-1213.83	-4287.67	-1819.55	-1754.24

Parameter estimates of the daily number of interactions. Recall that the the number of interactions is calculated as the sum of the daily number of retweets, quotes, likes and replies. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1%, respectively. In the lower rows the descriptive statistics of the standardized daily number of interactions for each company is provided, which directly enters the conditional volatility process. Lastly, the optimized Quasi Log Likelihood is given.

5.2.3 Model specification

Once the Twitter-based variables are computed, the lag order of the models is specified to estimate the parameters and construct the conditional volatility. More specifically, the lags P and Q must be determined for the mean model of equation (16), which is done via the Bayes Information Criterion, explained in equation (24). The selection of the lags in the mean model is done disregarding the specification of the conditional volatility, that is, the volatility is assumed to follow the benchmark model of equation (17). The combination of lags P and Q for $P, Q = 1, \ldots, 5$ that minimizes the BIC using the benchmark specification from equation (16) and equation (17) is selected for the conditional mean model. The lags P and Q are the same for every model in section 4.1. This procedure is repeated for each company from table 1.

The number of lags $R_k \geq 0$ for $k \in \{1, \ldots, n_K\}$ and $S_k \geq 0$ for $k \in \{1, \ldots, n_{K_1}\}$ for which the BIC is minimized is chosen as the optimal lag order. This search is conducted per company for each conditional volatility specification in equation (27). The lag order R_k (S_k) that minimizes the BIC in the constant parameter setting is subsequently used for both the constant parameter models as for the models that allow for varying parameters. As aforementioned, the BIC is more restrictive than for instance the AIC, and thus penalizes the inclusion of more parameters more severe.

Only for seven companies, the mean returns process follows a higher order lag autoregressive moving-average model. This is the case for Apple, Amazon, Microsoft, Netflix, NIKE, Tesla, and Walmart. With regard to the exogenous Twitter-based variables, it can be seen that the BIC successfully excluded redundant order lags. In the mean equation, only for Amazon, Exxon Mobil, Microsoft, and Netflix do longer lags of the Twitter-based variables explain the mean stock return.

For the models where the Twitter-based exogenous variables enter the volatility equation (model II &

model III), almost no companies include higher order lags of the exogenous variables. For model II, Microsoft and Tesla incorporate longer lags of the Twitter-based sentiment. This indicates that only for these companies, Twitter-based sentiment has effect on the conditional volatility for multiple periods. Apple and Amazon include multiple lags of the disagreement measure volatility of the sentiment and of the sentiment itself, respectively. Since using the BIC, for almost no companies higher order lags are included in the volatility models, it can be concluded that when Twitter effects the conditional volatility, Twitter can only explain volatility in the subsequent period.

Additionally, note that section 4.2 provides the assumptions that must be satisfied, such that a strictly stationary process σ_t exists. Hence, these assumptions are checked after the quasi log-likelihood is minimized[14], to see whether these hold. Assumption 1 and assumption 4 are imposed as constraints on the SLSQP minimization, which guarantees that these constraints are satisfied for all models. Note that for five companies, $P = Q = 1$, such that it must hold that $|\gamma_1| < 1$ and $|\delta_1| < 1$. Similarly, boundaries are provided for all elements in φ, which guarantees the non-negativity constraints and guarantees that φ is in the interior of the parameter space. To examine whether assumption 2, assumption 3, and assumption 5 hold, histograms of the realizations of the unobserved innovations $\tilde{z}_t = \frac{\tilde{\varepsilon}_t(\hat{\vartheta})}{\tilde{\sigma}_t(\hat{\theta})}$ and $\eta_t^{(x)} = \frac{\tilde{\epsilon}_t^{(x)}(\hat{\vartheta}_x)}{\tilde{\sigma}_t^{(x)}(\hat{\theta}_x)}$ are plotted after estimation. These histograms provide an intuition in the (conditional) moments of the innovations. Moreover, the first and second sample moments are calculated. These histograms and moments are in line with the assumptions. Lastly, note that the exogenous variables in $\boldsymbol{x}_t$ are checked to be stationary by means of the ADF test, and the null hypothesis is rejected at the 5% level for all instances, indicating that these variables all form a stationary process, thus satisfying assumption 2.

[14]Minimization is done using the *SLSQP* algorithm from *SciPy* in Python.

6 Results

This chapter discusses the results of the parameter estimation of the constant and varying parameter models. First, the parameter estimates of the volatility models under the constant and the varying parameter specification are presented and discussed. These estimates provide insights into the research question, which aims to answer if, how, and when Twitter can be leveraged to explain conditional volatility. Subsequently, a comparison between the various models is conducted to identify successful practices in including the exogenous variables, and to distinguish similarities across the selected companies. It is investigated whether allowing for varying parameters based on macroeconomic and market conditions can significantly improve volatility models. Finally, by using the squared residuals of the benchmark model as a proxy for the conditional volatility, the in-sample fit of the models is evaluated using the MSE and QLIKE loss functions.

6.1 Parameter estimates

This section discusses the parameter estimates of the proposed models. First, the parameters that describe the mean return series are analysed. Then, the volatility persistence under the inclusion of exogenous variables and varying parameters are compared with the benchmark model. Thereafter, the parameter estimates that include the Twitter-based variables are analysed to infer how these variables affect the mean return series and the volatility process of the stock price returns.

By minimizing the quasi log-likelihood from equation (28) for each conditional volatility specification in equation (27), the parameters estimates for each model are found. Similarly, the parameter estimates for the models that allow varying parameters in the conditional volatility equation are found as well as the argument $\hat{\varphi}(\boldsymbol{\zeta}_t)$ that minimizes $\mathcal{Q}_T$ from equation (28). Note that for the k-th element of $\hat{\varphi}(\boldsymbol{\zeta}_t)$ it holds that $\hat{\varphi}_k(\boldsymbol{\zeta}_t) = \hat{\varphi}_k + \hat{\varphi}_k^* G(\boldsymbol{\zeta}_t, 25, \bar{\boldsymbol{\zeta}})$. Using the t-test values calculated by equation (35), the significance of the individual parameter estimates is determined.

For the notation of the parameters, let $\pi_{metric,j}$ denote the parameter describing the effect of the lag j value of Twitter metric *metric*, where possible Twitter metrics are *sent*, *tweet*, and *interact*, describing the sentiment score, number of tweets and number of interactions, respectively. Similar notations hold for the parameters describing the asymmetric effects of the Twitter sentiment $\psi_{metric,j}$, and for the parameters describing the effect of the conditional volatility of the exogenous variables $\lambda_{metric,j}$.

6.1.1 ARMA process and volatility persistence

Table 7, table 8, and table 9 show the parameter estimates that minimize $\mathcal{Q}_T$ for the ARMAX-apARCH model, the ARMA-apARCH-apX model and the ARMA-apARCH-apXGARCH model, respectively. By inspecting the parameters that describe the conditional mean in either of these models, it can be seen that for Model I, for almost no companies the auto-regressive parameters nor the moving-average parameters are significantly different from zero. This implies that when Twitter-variables are used to model the mean return series, these variables convey more information than the auto-regressive or moving-average parts of the model. Furthermore, the magnitude of these parameters is particularly low when comparing them to the parameters modeling AR and MA part in Model II and Model III.

By considering the significant auto-regressive and moving-average parameters from Model II and Model III, it can be seen that the lagged value of the stock price return has a negative effect on current returns. From the generally positive estimated parameter value δ_1 it can be derived that large positive (negative) values of the unanticipated returns on the previous day signals higher (lower) returns in subsequent days. The sign and magnitude of the parameter estimates of the mean return equation shown in table 8 and table 9 are comparable with the parameter estimates of the benchmark model in table 14 in appendix A.1.3, although the signs are opposite for Netflix and salesforce.

Table 15, table 16, and table 17 in appendix A.1.4 show the parameter estimates for the models that allow for varying parameters in the conditional volatility equation. The parameters including the AR and MA parts in the conditional mean are similar to the parameter estimates observed in their constant parameter counterparts with respect to magnitude and significance. Considering the significance of the AR and MA parameters at at least the 10% level across all companies for all models (excluding for model I), it can be concluded that the auto-regressive moving-average specification properly explain the mean return series. Moreover, this supports the remark from Franq and Zakoian (2004) that a pure GARCH specification is too restrictive in financial time series data.

Table 7 – Parameter estimates of Model I with constant parameters (data ranges from 01-01-2011 until 31-08-2021).

	Parameter estimates per company of ARMAX-apARCH model											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
ARMAX parameters												
μ	0.162***	0.243***	-0.02	0.024***	-0.016	0.073***	0.189***	0.015	0.151***	0.0	0.278**	0.068**
	(4.1e-02)	(6.9e-02)	(4.3e-02)	(9.7e-03)	(9.0e-02)	(3.1e-02)	(5.9e-02)	(3.3e+00)	(4.9e-02)	(5.9e-02)	(1.2e-01)	(3.0e-02)
γ_1	0.025	0.009	0.015	0.006	-0.023	0.011	0.012*	-0.022	0.03	0.003*	0.027	0.002
	(2.9e-02)	(3.4e-02)	(1.2e-02)	(1.6e-02)	(2.6e-02)	(1.8e-02)	(9.2e-03)	(1.1e+00)	(2.8e-02)	(2.2e-03)	(3.1e-02)	(2.0e-02)
γ_2		-0.069*						0.018	-0.03			0.025
		(4.3e-02)						(2.6e-01)	(3.8e-02)			(2.4e-02)
γ_3									0.08***			
									(2.8e-02)			
δ_1	-0.051*	-0.031	0.014	0.016	-0.038*	0.022**	0.025**	-0.005	-0.683***	0.0	-0.055*	0.034**
	(3.3e-02)	(3.5e-02)	(1.7e-02)	(2.1e-02)	(2.6e-02)	(1.3e-02)	(1.3e-02)	(1.0e-02)	(1.7e-02)	(3.4e-04)	(3.5e-02)	(1.5e-02)
δ_2	-0.032						-0.029***	-0.036			0.024	
	(4.8e-02)						(9.3e-03)	(1.0e+01)			(3.3e-02)	
δ_3	-0.38***											
	(2.6e-02)											
$\pi_{sent,1}$	0.403***	0.007	-0.033	0.014	-0.006	-0.009	0.009*	0.018	-0.08***	-0.001	-0.918***	-0.443***
	(7.6e-06)	(2.4e-02)	(2.6e-02)	(1.9e-02)	(2.6e-02)	(1.3e-02)	(5.2e-03)	(7.2e-02)	(2.8e-02)	(2.1e-01)	(3.4e-03)	(6.3e-02)
$\pi_{sent,2}$		-0.745***			-0.026		-0.979***	0.021				
		(1.9e-02)			(2.5e-02)		(2.1e-03)	(1.1e-01)				
$\pi_{tweets,1}$	0.016	-0.007	-0.92***	-0.588***	-0.864***	-0.694***	0.933***	-0.029	-0.055**	0.988***	0.914***	-0.084***
	(2.2e-02)	(2.1e-02)	(2.0e-03)	(1.1e-01)	(5.6e-03)	(3.6e-02)	(1.8e-03)	(1.8e+00)	(2.5e-02)	(4.7e-03)	(2.9e-04)	(2.9e-02)
$\pi_{tweets,2}$								0.981***				
								(5.5e-02)				
$\pi_{interact,1}$	-0.043**	0.747***	0.909***	0.567***	0.858***	0.684***	-0.059***	0.01	0.636***	-1.003	-0.003	0.386***
	(2.4e-02)	(9.3e-04)	(1.7e-04)	(1.6e-03)	(1.8e-04)	(1.3e-04)	(5.6e-05)	(2.6e-01)	(9.6e-05)	(5.5e+00)	(8.3e-03)	(2.1e-05)
$\pi_{interact,2}$								-0.999				
								(2.4e+00)				
apARCH parameters												
ω	0.193***	0.391***	0.048***	0.082***	0.036***	0.138***	0.332***	0.028***	0.273***	0.04***	0.224***	0.377***
	(9.3e-06)	(6.4e-06)	(9.4e-08)	(4.3e-07)	(1.0e-60)	(2.6e-26)	(3.7e-05)	(6.9e-03)	(5.3e-25)	(2.4e-03)	(5.4e-06)	(9.5e-06)
β	0.814***	0.763***	0.894***	0.84***	0.917***	0.763***	0.684***	0.92***	0.797***	0.899*	0.942***	0.511***
	(4.5e-04)	(8.4e-04)	(1.7e-05)	(9.6e-04)	(6.0e-06)	(3.1e-05)	(7.7e-04)	(1.2e-02)	(6.3e-05)	(6.5e-01)	(1.2e-04)	(6.6e-04)
α	0.106***	0.145***	0.078***	0.074***	0.033***	0.06***	0.178***	0.075***	0.053***	0.091***	0.039***	0.244***
	(1.1e-03)	(4.4e-03)	(4.5e-04)	(7.3e-03)	(3.5e-04)	(3.5e-03)	(1.6e-03)	(1.3e-09)	(3.4e-03)	(4.4e-07)	(4.0e-04)	(1.6e-02)
ψ	0.463***	0.19*	0.35***	0.336***	1.0*	1.0***	0.208**	0.236	1.0***	0.333	-0.0	0.052
	(8.8e-02)	(1.2e-01)	(7.5e-02)	(1.3e-01)	(7.1e-01)	(2.5e-01)	(9.4e-02)	(2.5e+01)	(2.6e-01)	(2.3e+00)	(8.1e-02)	(1.2e-01)

Parameter estimates per company of the ARMAX-apARCH model proposed in equation (18). Table indicates the parameters of the conditional mean equation and the conditional variance equation. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1% significance level, respectively. In case the standard error is infinitely small, it is manually set to 1.0e-60, and the corresponding parameter is denoted significant at the 1% level. Number of lags of each variable that enters the model is determined by the BIC.

Subsequently, the parameters that describe the conditional volatility process are considered. The constant parameter estimates show that the intercept in the volatility models, ω is reduced by a substantial margin when exogenous variables enter the conditional volatility equation. This reduction is observed even more when the parameters are allowed to vary as a function of the VIX and the TED rate, shown by the tables in appendix A.1.4. This implies that the volatility models are more flexible under the inclusion of exogenous variables and/or varying parameters.

From table 7, table 8, and table 9 it is deduced that for all companies an apARCH process is present in the conditional volatility. With the exception of Tesla, the parameter estimates for all companies show that the estimates for α and β are significantly different from zero at at least the 10% level. The magnitude of β implies that shocks to the volatility process decay slowly over time. These shocks are caused by either unexpected values of the mean return process (via α), or via large values of the exogenous variables (for model II and model III). Note that in the constant parameter setting, the magnitude of the estimates for α is rather small, indicating that the immediate effect of shocks is small, although shocks affect the volatility for a long period. Moreover, throughout all constant parameter models, almost all the selected companies experience asymmetric effects of unexpected returns, which supports the leverage effect observed by Black (1976). Only for the conditional volatility of Tesla, the asymmetric effect is estimated to be negative, though this estimate is not significantly different from zero.

The varying parameter estimates, which are tabulated in appendix A.1.4, show substantial differences compared to the constant parameter estimates. The magnitude of the estimated parameter β^* is substantial for McDonalds, Microsoft, Netflix, NIKE, salesforce, and Walmart, which implies that the effects of σ_{t-1}^2 on the volatility are largely dependent on $\boldsymbol{\zeta}_t$. For Netflix, NIKE, and salesforce, the parameter $\beta(\boldsymbol{\zeta}_t)$ increases when $G(\cdot) = 1$ across all models, and for the other companies, $\beta(\boldsymbol{\zeta}_t)$ decreases under $G(\cdot) = 1$. Additionally, $\alpha^* > 0$ for almost all companies, which indicates that during economic downturn, shocks ε_{t-1}^2 to the mean return series explains a larger part of the volatility than during times of economic prosperity.

For the majority of the companies, the estimates for ψ and ψ^* again show sizable evidence in favor of the leverage effect. However, the parameter estimates for ψ are less significant compared to the constant parameter setting. For Netflix, it is actually the case that positive unexpected returns have more effect on the volatility than negative returns of the same magnitude under $G(\cdot) = 1$. For Coca-Cola and McDonalds, the magnitude of $\psi(\boldsymbol{\zeta}_t)$ significantly decreases under economic downturn, which indicates that the asymmetric effect decreases in magnitude.

Both in the constant parameter setting, as well as in the varying parameter setting, it is noted that across multiple models, ψ approaches one in the case for Coca-Cola, McDonalds, NIKE, and salesforce. As aforementioned, this indicates that positive unexpected returns have negligible effect on the volatility series.

From the parameter estimates, it can be seen that when exogenous variables enter the mean return series or the conditional volatility series, persistence (measured by $\alpha(1 + \psi^2) + \beta$) generally increases. This indicates that when Twitter-based variables enter the model, shocks to the volatility process take longer to decay over time. In contrast to the results by Kim and Kon (1999) and Lamoureux and Lastrapes (1990), accommodating for possible structural breaks in the volatility process does not reduce the degree of persistence. Although persistence generally decreases during economic prosperity, persistence during economic downturn is actually higher compared to persistence in the constant parameter case. This is observed for all companies, except McDonalds and Walmart. Hence, under $G(\cdot) = 1$, shocks to the conditional variance generally 'persist' for longer amounts of time, which is consistent with the intuition that during economic downturn, high levels of volatility indicate increased volatility in subsequent periods. Similarly to the constant parameter case, it can be concluded that the conditional volatility series exhibit a high degree of persistence.

Table 8 – Parameter estimates of Model II with constant parameters (data ranges from 01-01-2011 until 31-08-2021).

	Parameter estimates per company of ARMA-apARCH-apX model											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
ARMA parameters												
μ	0.17***	0.244*	-0.015	0.024	0.003	0.079***	0.19***	0.3**	0.141**	0.003	0.576	0.064*
	(4.1e-02)	(1.7e-01)	(5.2e-02)	(2.9e-02)	(4.2e-02)	(3.2e-02)	(4.2e-02)	(1.5e-01)	(8.3e-02)	(3.4e-03)	(1.1e+02)	(4.0e-02)
γ_1	-0.419***	-0.724	-0.874**	-0.589***	-0.708***	-0.694***	-0.982***	-0.494***	-0.685***	0.965***	-1.0	-0.617***
	(3.5e-02)	(6.9e-01)	(4.7e-01)	(4.1e-02)	(9.4e-02)	(5.6e-02)	(4.8e-02)	(2.9e-02)	(1.4e-01)	(2.2e-03)	(1.6e+02)	(1.1e-01)
γ_2		-0.006						-0.649***	-0.077			-0.05
		(2.5e-01)						(5.6e-02)	(9.7e-02)			(5.1e-02)
γ_3									-0.055			
									(5.0e-02)			
δ_1	0.441***	0.725***	0.858***	0.568***	0.688***	0.686***	0.935***	0.496***	0.64***	-0.987***	1.0	0.566***
	(1.7e-05)	(5.6e-02)	(1.3e-02)	(1.7e-03)	(7.9e-03)	(9.2e-03)	(1.2e-01)	(2.2e-02)	(6.3e-03)	(9.9e-03)	(8.5e+00)	(1.4e-03)
δ_2	0.017						-0.055***	0.628***			-0.004	
	(2.4e-02)						(2.1e-04)	(1.1e-03)			(1.5e-02)	
δ_3	-0.04**											
	(2.3e-02)											
apARCH-apX parameters												
ω	0.098***	0.13***	0.044***	0.078***	0.035***	0.147***	0.232***	0.038***	0.245***	0.134***	0.229***	0.077***
	(1.0e-60)	(1.0e-60)	(1.0e-60)	(1.0e-60)	(1.0e-60)	(3.6e-27)	(1.0e-60)	(2.8e-28)	(6.1e-29)	(1.2e-27)	(1.0e-60)	(1.0e-60)
β	0.818***	0.818***	0.899***	0.844***	0.882***	0.726***	0.709***	0.976***	0.788***	0.787***	0.921	0.878***
	(3.8e-04)	(6.1e-02)	(1.2e-02)	(1.0e-03)	(1.0e-02)	(9.8e-03)	(9.1e-02)	(7.4e-04)	(7.9e-03)	(7.5e-03)	(3.0e+00)	(1.2e-03)
α	0.099***	0.106***	0.071***	0.073***	0.099***	0.113***	0.157***	0.007***	0.055***	0.06***	0.049***	0.029***
	(1.0e-60)	(1.0e-60)	(3.6e-25)	(1.0e-60)	(1.0e-60)	(2.4e-24)	(1.0e-60)	(6.8e-09)	(1.0e-60)	(2.5e-08)	(9.0e-24)	(3.7e-20)
ψ	0.472***	0.177	0.369***	0.333***	0.153***	0.528***	0.178**	1.0***	0.999**	1.0*	-0.014	1.0***
	(9.0e-02)	(2.4e-01)	(1.4e-01)	(1.3e-01)	(5.3e-02)	(1.2e-01)	(9.6e-02)	(5.4e-02)	(5.8e-01)	(7.7e-01)	(1.3e+01)	(5.6e-02)
$\pi_{sent,1}$	0.015***	0.0	0.0	0.0***	0.0	0.002	0.02	0.0	0.0	0.129	0.026	0.007
	(2.2e-03)	(6.7e-10)	(2.1e-09)	(2.2e-11)	(1.1e-05)	(1.7e-03)	(1.9e-01)	(1.9e-09)	(1.1e-10)	(1.3e-01)	(6.6e+01)	(6.9e-03)
$\pi_{sent,2}$							0.038				0.0	
							(1.9e-01)				(2.1e-06)	
$\pi_{tweets,1}$	0.079***	0.122	0.001	0.0	0.0	0.011	0.0*	0.21***	0.124	0.0	0.174	0.005
	(7.7e-03)	(1.2e-01)	(9.3e-04)	(3.0e-04)	(1.4e-03)	(3.1e-02)	(7.8e-11)	(1.9e-03)	(1.3e-01)	(2.6e-10)	(2.0e+01)	(3.4e-02)
$\pi_{interact,1}$	0.165**	1.15*	0.014	0.001***	0.001***	0.0***	0.076	0.004***	0.007	0.722	0.0	0.014
	(7.2e-02)	(7.5e-01)	(3.2e-02)	(9.5e-05)	(1.8e-04)	(2.4e-09)	(3.8e-01)	(1.2e-04)	(3.6e+00)	(1.2e+00)	(1.4e-07)	(2.0e-02)
$\psi_{sent,1}$	0.998***	0.807	0.151	-1.0**	-0.976	0.999	-0.992	0.852	-0.62	0.348**	0.999	-1.0
	(3.8e-01)	(6.3e+00)	(8.0e+00)	(5.4e-01)	(1.0e+03)	(1.6e+00)	(3.7e+00)	(2.2e+01)	(1.7e+00)	(1.9e-01)	(2.1e+03)	(1.1e+00)
$\psi_{sent,2}$							-0.375				0.884	
							(7.8e-01)				(3.4e+03)	
$\psi_{tweets,1}$	0.447*	0.292*	0.033	-0.797***	-0.247	1.0***	0.44*	0.916***	0.727**	-0.484	1.0	1.0*
	(2.9e-01)	(2.1e-01)	(4.8e-01)	(2.3e-01)	(1.9e+00)	(1.0e-01)	(2.7e-01)	(1.6e-01)	(3.4e-01)	(4.3e+00)	(6.4e+01)	(6.1e-01)
$\psi_{interact,1}$	0.694***	0.974	0.538	0.244	-0.509	-0.915**	1.0***	0.231	0.995	0.57	1.0	1.0***
	(2.8e-01)	(1.3e+00)	(1.7e+00)	(8.0e-01)	(9.9e-01)	(5.3e-01)	(3.8e-01)	(3.7e-01)	(3.2e+00)	(1.1e+00)	(1.7e+02)	(1.3e-01)

Parameter estimates per company of the ARMA-apARCH-apX model proposed in equation (20). Table indicates the parameters of the conditional mean equation and the conditional variance equation. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1% significance level, respectively. In case the standard error is infinitely small, it is manually set to 1.0e-60, and the corresponding parameter is denoted significant at the 1% level. Number of lags of each variable that enters the model is determined by the BIC.

6.1.2 Significance of Twitter-based sentiment variables

To determine how Twitter exerts influence on the volatility of the stock price returns, the parameters that include the Twitter-based exogenous variables are inspected. This subsection provides a detailed description of the parameter estimates and discusses the implied effects of Twitter on stock price return volatility.

Table 9 – Parameter estimates of Model III with constant parameters (data ranges from 01-01-2011 until 31-08-2021).

	Parameter estimates per company of ARMA-apARCH-apXGARCH model											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
ARMA parameters												
μ	0.164***	0.236	-0.014	0.024	0.004	0.076**	0.18	0.304***	0.147	0.151**	0.582	0.068**
	(4.1e-02)	(1.9e-01)	(4.3e-02)	(3.1e-02)	(5.4e-02)	(3.6e-02)	(1.9e-01)	(1.2e-01)	(2.9e-01)	(6.8e-02)	(1.2e+02)	(3.4e-02)
γ_1	-0.4***	-0.666**	-0.876**	-0.596***	-0.707***	-0.711***	-0.958***	-0.839***	-0.689	-0.732***	-1.0	-0.514***
	(3.3e-02)	(3.9e-01)	(3.9e-01)	(4.1e-02)	(9.5e-02)	(1.4e-01)	(8.8e-02)	(1.4e-01)	(7.5e-01)	(6.3e-02)	(1.7e+02)	(5.3e-02)
γ_2		-0.003						0.069**	-0.094			-0.082***
		(1.7e-01)						(3.5e-02)	(5.3e-01)			(3.3e-02)
γ_3									-0.065			
									(2.6e-01)			
δ_1	0.422***	0.667***	0.861***	0.576***	0.687***	0.705***	0.912***	0.85***	0.641***	0.706***	1.0	0.458***
	(1.2e-05)	(1.0e-01)	(3.7e-03)	(2.3e-03)	(1.2e-02)	(2.9e-02)	(9.6e-02)	(2.3e-03)	(1.4e-02)	(1.1e-02)	(2.0e+01)	(2.4e-05)
δ_2	0.016						-0.049***	-0.057***			-0.004	
	(2.3e-02)						(6.9e-04)	(5.7e-05)			(6.1e-02)	
δ_3	-0.038*											
	(2.4e-02)											
apARCH-apXGARCH parameters												
ω	0.13***	0.282***	0.045***	0.069***	0.036***	0.089***	0.001***	0.173***	0.208***	0.324***	0.377***	0.187***
	(1.0e-60)	(1.6e-25)	(1.0e-60)	(1.0e-60)	(1.0e-60)	(1.0e-60)	(4.8e-28)	(1.8e-27)	(1.0e-60)	(3.5e-24)	(2.1e-18)	(3.5e-28)
β	0.797***	0.804***	0.898***	0.849***	0.883***	0.705***	0.716***	0.959***	0.771***	0.774***	0.919	0.516***
	(2.6e-04)	(1.2e-01)	(2.8e-03)	(3.4e-03)	(1.6e-02)	(2.9e-02)	(7.6e-02)	(2.2e-03)	(1.5e-02)	(1.3e-02)	(1.3e+01)	(1.4e-04)
α	0.097***	0.126***	0.07***	0.071***	0.098***	0.118***	0.157***	0.023***	0.06***	0.105***	0.047***	0.231***
	(1.7e-26)	(1.0e-60)	(1.6e-31)	(1.1e-33)	(3.9e-26)	(5.9e-37)	(1.0e-60)	(1.0e-60)	(1.0e-60)	(1.1e-26)	(1.0e-60)	(2.2e-34)
ψ	0.556***	0.132	0.363***	0.34***	0.154*	0.543***	0.211**	0.375***	1.0	0.579***	-0.009	0.052
	(9.8e-02)	(1.5e-01)	(8.7e-02)	(1.1e-01)	(1.2e-01)	(1.2e-01)	(1.2e-01)	(1.4e-01)	(3.1e+00)	(9.5e-02)	(4.9e+01)	(1.2e-01)
$\pi_{sent,1}$	0.011	0.0	0.0**	0.0***	0.311***	0.0	0.0	0.0	0.0	0.024***	0.0	0.0***
	(2.2e-02)	(2.6e-09)	(5.9e-11)	(8.5e-10)	(5.0e-02)	(3.2e-09)	(1.9e-08)	(2.9e-10)	(1.6e-04)	(1.6e-03)	(9.1e-07)	(2.7e-11)
$\pi_{sent,2}$		0.0										
		(1.1e-10)										
$\pi_{tweets,1}$	0.0	0.002***	0.0	0.0***	0.015*	0.0***	0.0	0.0**	0.001	0.0***	0.005	0.0
	(3.3e-05)	(1.8e-04)	(2.4e-07)	(3.1e-11)	(1.1e-02)	(2.1e-11)	(2.0e-10)	(1.0e-09)	(5.8e-03)	(2.9e-11)	(2.5e-01)	(7.9e-09)
$\pi_{interact,1}$	0.0	0.0	0.0*	0.0	0.0	0.0***	0.0***	0.0	0.0	0.015*	0.0	0.0
	(1.1e-04)	(2.5e-05)	(4.4e-07)	(2.0e-07)	(4.0e-06)	(2.2e-11)	(1.5e-09)	(1.8e-07)	(1.7e-09)	(9.8e-03)	(9.3e-08)	(6.1e-06)
$\lambda_{sent,1}$	0.0	0.0	0.0	0.0	0.0	0.085***	0.252***	0.0	0.0	0.0	0.0	0.181***
	(1.8e-07)	(4.0e-06)	(1.2e-05)	(9.4e-03)	(1.7e-04)	(3.5e-03)	(1.1e-02)	(2.5e-06)	(1.3e-05)	(3.0e-06)	(2.5e-03)	(4.3e-03)
$\lambda_{sent,2}$	0.0											
	(3.3e-07)											
$\lambda_{sent,3}$	0.0											
	(4.8e-06)											
$\lambda_{tweets,1}$	0.067***	0.0	0.0	0.01***	0.0	0.0	0.069***	0.0	0.05	0.0	0.0	0.023
	(7.6e-03)	(4.0e-05)	(6.7e-06)	(2.0e-03)	(2.6e-04)	(5.1e-06)	(2.4e-02)	(4.2e-02)	(3.4e-01)	(3.4e-06)	(1.4e-03)	(7.0e-02)
$\lambda_{interact,1}$	0.062***	0.0	0.011**	0.0	0.0	0.0	0.0	0.0	0.087***	0.0	0.018	0.0
	(8.4e-03)	(7.4e-05)	(5.7e-03)	(4.9e-03)	(2.6e-04)	(3.3e-06)	(2.7e-02)	(2.9e-06)	(8.7e-03)	(8.1e-07)	(1.1e+02)	(1.1e-01)
$\psi_{sent,1}$	0.594	0.816	-0.258	-1.0**	1.0***	0.054	0.563	0.249	1.0***	-0.259	0.289	-0.187
	(4.9e-01)	(3.1e+00)	(9.2e-01)	(5.1e-01)	(7.2e-08)	(2.1e+01)	(1.2e+02)	(8.3e-01)	(0.0e+00)	(3.1e-01)	(3.4e+03)	(3.1e-01)
$\psi_{sent,2}$		-0.66										
		(4.2e+00)										
$\psi_{tweets,1}$	0.972***	0.333	-0.971	0.104	0.986***	0.329***	0.073	0.497**	0.753	-0.715	1.0***	0.106
	(9.2e-03)	(7.6e-01)	(3.4e+00)	(2.8e-01)	(9.5e-02)	(1.3e-01)	(3.8e-01)	(2.9e-01)	(1.2e+00)	(6.0e-01)	(6.8e-07)	(5.1e-01)
$\psi_{interact,1}$	0.732	0.501	0.007	0.075	0.38	0.165	0.266*	0.095	0.282	0.681	0.671	0.174
	(1.4e+00)	(6.3e+00)	(7.6e-01)	(1.2e+00)	(4.8e+00)	(1.9e-01)	(1.7e-01)	(1.1e+00)	(8.9e+00)	(1.2e+00)	(3.2e+02)	(5.0e+00)

Parameter estimates per company of the ARMA-apARCH-apXGARCH model proposed in equation (23). Table indicates the parameters of the conditional mean equation and the conditional variance equation. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1% significance level, respectively. In case the standard error is infinitely small, it is manually set to 1.0e-60, and the corresponding parameter is denoted significant at the 1% level. Number of lags of each variable that enters the model is determined by the BIC.

Before the behaviour of the effect of the Twitter-based exogenous variables on the conditional volatility is analysed, the first consideration is whether the inclusion of these variables into the mean return process provides reasonable estimates. From table 7, it is seen that for six companies, the parameter $\pi_{sent,1}$ is significant, which indicates that the lagged value of the Twitter sentiment can explain the mean stock price return. For Microsoft and Amazon, the second order lag of the sentiment has effect on the conditional mean of the stock returns. With the exception of Apple, all significant parameter estimates indicate a negative effect of sentiment on the stock price return. This means that higher sentiment scores explains lower returns in the subsequent period, and vice versa. The effect of the daily number of tweets and the daily number of interactions on the conditional return process are significant across more companies. Chevron, Coca-Cola, Exxon Mobil, McDonalds, NIKE, and Walmart all show significant negative parameter values of $\pi_{tweets,1}$. Interestingly, Microsoft, salesforce and Tesla report significant positive coefficients that describe the effect of the number of tweets on returns. With the exception of Netflix, salesforce and Tesla, the parameter estimate $\pi_{interact,1}$ is significantly different from zero at at least the 5% level. For Apple and Microsoft this effect is negative, which signals that a high number of interactions signals lower stock price returns the next day, and vice versa.

Comparison of the results of the constant parameter estimates of model I from table 7 with their varying-parameter counterparts, reported in table 15 in appendix A.1.4, reveals that the parameter estimates for the Twitter-based sentiment variables are generally equal, both in sign, significance and magnitude.

Combining the significance of the Twitter-based variables on the mean return series with the low magnitude of the auto-regressive moving-average parameters in these models, the following can be concluded. For almost all companies, regardless of the state of the economy, the Twitter-based variables provide reasonable explanatory power for the mean return series of the companies in this paper.

While it can be concluded that Twitter-based variables can successfully explain the mean return series, the main question of this paper addresses how these variables affect the conditional volatility of stock price returns. The parameter estimates of the constant parameter specification of model II and model III are given in table 8 and table 9, respectively. The varying parameter estimates of these models can be found in table 16 and table 17 in appendix A.1.4.

From table 8, it can be deduced that only in the case of Apple, the parameters that describe the effect of the Twitter-based variables are all significantly different from zero. Additionally, $\psi_{sent,1}$ approaches one, which implies that the effect of positive Twitter sentiment on conditional volatility is negligible. For Netflix, both the standardised number of tweets and the standardised number of interactions successfully explain the conditional volatility. Finally, the effect of the standardised number of interactions is significant at at least the 10% significance level for Amazon, Coca-Cola, Exxon Mobil, and Netflix. Hence, via the specification of model II, Twitter-based variables successfully explain the volatility of only five companies. This indicates that in general, this model lags flexibility, or outliers in the time series of the exogenous variables do not correlate with the stock price return volatility.

Therefore, it is investigated whether the volatility of these exogenous variables significantly effects the conditional volatility process. Since the conditional volatility $\sigma_t^{(x)}$ is positive and stationary by definition, the volatility of the exogenous variables can directly enter the conditional volatility equation of the stock price returns. These processes serve as a disagreement measure, that capture the degree of variation to which the exogenous variables vary over time. Recall that based on the time series plots in appendix A.2.1, it is expected that this specification may be particularly beneficial for Chevron, Exxon Mobil, McDonalds, salesforce, Tesla, and Walmart.

The parameter estimates of the model of equation (23) are substantially different than the estimates of model II. For McDonalds, Microsoft, and Walmart, the estimate of $\lambda_{sent,1}$ is significantly different from zero. This implies that the conditional volatility of the Twitter-based sentiment can explain volatility of stock price returns. Increased variation of Twitter-based sentiment can thus be used to account for increased conditional volatility for these companies. Moreover, $\lambda_{tweets,1}$ and $\lambda_{interact,1}$ determine the effect of the conditional volatility of the standardised number of tweets and the standardised number of interactions, respectively. Table 9 shows that volatility of the daily number of tweets has a significant effect on conditional volatility in the subsequent period for Apple, Coca-Cola and Microsoft. Similarly, $\lambda_{interact,1}$ is significant for Apple, Chevron, and NIKE. This finding is in contrast with the expected effects of the volatility of the exogenous based on the plots in appendix A.2.1. Particularly in the case of Apple and Microsoft, the plots show little indication that these variables have significant effect on the conditional volatility, since these plots display little variation of the volatility of the exogenous variables.

Additionally, the Twitter-based variables can successfully explain the volatility series for three more companies. For Amazon, $\pi_{tweets,1}$ is significantly different from zero, for Exxon Mobil both $\pi_{sent,1}$ and $\pi_{tweets,1}$ are significantly different from zero. Furthermore, notice that $\psi_{sent,1} \approx 1$, which indicates that only negative values of Twitter sentiment significantly effects the conditional volatility of Exxon Mobil. For salesforce, both $\pi_{sent,1}$ and $\pi_{interact,1}$ are significantly different from zero.

In total, for all companies except Netflix and Tesla, it holds that either λ_k or π_k is significantly different from zero for at least one exogenous variable k. Note this was only that case for five companies by the constant parameter estimates of model II. Hence, it can be concluded that for the majority of the companies, the volatility of Twitter-based variables successfully explain the conditional volatility of stock price returns. Especially for Chevron, Exxon Mobil, McDonalds, Microsoft, NIKE, salesforce, and Walmart, it is argued that the conditional volatility of the stock price returns is better described by model III than model II.

Additionally, it is considered whether the exogenous variables asymmetrically exert influence on the conditional volatility of stock price returns. Based on the parameter estimates in table 8 and table 9, the Twitter-based sentiment is only significant for Apple in model II, and for Exxon Mobil and salesforce in model III. As aforementioned, for Apple and Exxon Mobil, $\psi_{sent,1}$ is significant, and approaches one, which implies that only negative values of Twitter sentiment significantly explain the volatility of stock price returns for these companies. In line with the results of Audrino et al. (2020) it is expected that high values of the *attention variables* (*i.e.*, the standardised values of the number of tweets and interactions) have more effect on the conditional volatility. Recall that for $\psi_k < 0$, positive values of exogenous variable k have a larger effect on the conditional volatility than negative values of the same magnitude, and vice versa for $\psi_k > 0$. However, since table 5 and table 6 show that the magnitude of negative values is very low (due to the skewed distribution), the estimated values of ψ_k do not necessarily convey any information about the asymmetric effects of the attention variables. Only when ψ_k approaches (minus) one, these parameters convey information, indicating that (negative) positive values have negligible effect on the conditional volatility. Although the magnitude of ψ_k is generally high, it does not approach one in any case.

Following these parameter estimates, the following conclusions can be reached. Via the specification of equation (20), the Twitter variables generally fail to explain spikes in the conditional volatility series. Furthermore, the number of tweets is reported to convey the most information, and is reported to be significantly different from zero for five companies. When the parameter estimates from model III in table 9 are considered, it is seen that in generally, the Twitter-based variables significantly explain stock price return volatility across twice as many companies. This indicates that the volatility of the Twitter-based variables can provide useful to explain the conditional volatility. Furthermore, when the effect of Twitter-based sentiment is significant, it is found that negative sentiment is more useful to explain volatility, which is consistent with the literature (M. P. Chen et al., 2013; Smales, 2015). Lastly, no higher order lags are reported to be significant, indicating that Twitter only effects one day ahead stock price return volatility.

Subsequently, the parameter estimates of the varying parameter estimates of these models are considered. These estimates are tabulated in appendix A.1.4. As aforementioned, it is expected that during periods where $G(\cdot) = 1$, the effect of Twitter on the conditional volatility series is larger. Additionally, it is expected that during these periods, the asymmetric effects are more evident, signalling larger effects of negative Twitter-based sentiment (Hsu et al., 2021; Smales, 2014).

From table 16 in appendix A.1.4, it is evident that the varying parameters allow for a much more flexible relationship between Twitter and the conditional volatility. When $G(\cdot) = 0$, $\pi_{sent,1}$ is significantly different from zero for Microsoft, salesforce, Tesla, and Walmart. This effect is reduced (to zero) in economic downturn, albeit $\pi^*_{sent,1}$ is not significantly different from zero for all companies. Contrarily, in economic downturn, the conditional volatility of Apple and NIKE is significantly affected by the Twitter-based sentiment towards these companies. The estimates for $\psi_{sent,1}$ and $\psi^*_{sent,1}$ additionally provide evidence towards the intuition that negative sentiment has more effect on the conditional volatility, although this does not hold for Microsoft. Furthermore, during economic downturn, the second order lags of the Twitter-based sentiment exerts significant influence on the stock price return volatility of Microsoft and Tesla. For Microsoft, negative values of the second order lag of Twitter-based sentiment has more effect on the conditional volatility than positive values.

Contrary to the constant parameter estimates of model II, the standardised number of tweets significantly affects the conditional volatility series of all companies, excluding salesforce. For Apple, Amazon, Chevron, Exxon Mobil, Netflix, and Walmart, the magnitude of this effect is the largest when $G(\cdot) = 0$. However, for Netflix, $\psi_{tweet,1} \approx 1$, indicating that the magnitude of both positive and negative (negative values are very small) effects is negligible. During economic downturn, the effects of the standardised number of tweets significantly reduces (to zero) for Apple, Chevron, Exxon Mobil, and Netflix. On the contrary, Coca-Cola, McDonalds, Microsoft, NIKE, and Tesla experience increasing significant effects of the standardised number of tweets in economic downturn.

Similar to the standardised number of tweets, the standardised number of interactions also exerts influence on the majority of the companies. For Apple, salesforce, and Tesla, the standardised number of interactions can successfully explain the conditional volatility of stock price returns during economic prosperity, this effect is reduced in economic downturn. For Chevron, Coca-Cola, and Exxon Mobil, $\pi^*_{interact,1} > 0$ significantly, which implies that the effects of the standardised number of interactions increases in economic downturn. Comparable to the constant parameter specification, the standardised number of interactions can significantly explain the conditional volatility of Amazon, Netflix, and salesforce (for salesforce the parameter estimate was not significant). However, for Amazon $\psi_{sent,1} \approx 1$, and for Exxon Mobil $\psi_{interact,1} + \psi^*_{interact,1} \approx 1$, which are both significant, which implies that the effects of the standardised number of interactions on the conditional volatility are negligible.

The parameter estimates form the varying parameter specification of model III are given in table 17 in appendix A.1.4. Again, the volatility of stock price returns of McDonalds, Microsoft, and Walmart can be significantly explained by the volatility of the Twitter-based sentiment. For McDonalds and Walmart this holds when $G(\cdot) = 0$, and for Walmart, the effect significantly reduces to zero in economic downturn. For Microsoft, $\lambda_{sent,1}$ is only significantly different from zero when $G(\cdot) = 1$. This implies that the effect of the volatility of Twitter sentiment is dependent on the VIX and the TED rate.

Similarly, the volatility of the number of tweets significantly exerts influence on the conditional volatility of Apple, Coca-Cola, and Walmart when $G(\cdot) = 1$, and on Microsoft, NIKE, and Walmart when $G(\cdot) = 1$. Whereas this effect was only significant for Apple, Coca-Cola and Microsoft in the constant parameter case. When the parameters vary through time via the function from equation (34), $\lambda^*_{interact,1}$ is significant for six companies. This indicates that under $G(\cdot) = 1$, this variable explains the volatility series of three more companies (Coca-Cola, Microsoft, and Walmart) than under the constant parameter case. Regardless of the state of the economy, this variable successfully explains the conditional volatility of NIKE.

Compared to the constant parameter case, the parameter estimates $\pi_{sent,1}$, $\pi_{tweets,1}$ and $\pi_{interact,1}$ are significant for more companies. During economic downturn, Twitter-based sentiment significantly explains the conditional volatility process of Apple, Chevron, Microsoft, Netflix, NIKE, and salesforce. Similar to the constant parameter case, $\pi_{sent,1}$ is significant for Exxon Mobil and salesforce. Additionally, negative sentiment has a larger effect on the volatility series than positive sentiment, and consistent with Smales (2014), these asymmetric effects increase during economic downturn for Coca-Cola and NIKE.

The standardised number of tweets exerts significant effects on the volatility in at least one state of the economy for eight companies, compared to the mere two times when the parameters are assumed to be constant. During economic prosperity, the parameter estimate of $\pi_{tweets,1}$ is significantly different from zero only for Apple, Amazon, and Cehvron. Yet, under $G(\cdot) = 1$, the standardised number of tweets successfully explains the conditional volatility of Coca-Cola, Microsoft, Netflix, NIKE, and Walmart. For Coca-Cola en Exxon Mobil, the parameter estimate $\pi^*_{interact,1}$ is significantly different from zero as well, which indicates that in periods of economic downturn, the standardised number of interactions significantly influences the conditional volatility.

In line with the proposed assumptions, allowing for varying parameters provides a more flexible modeling structure, and the significant effect of Twitter-based variables on stock price returns volatility is perceived more often. Compared to the constant parameter estimates of model II, Twitter-based variables significantly explain the conditional volatility of seven more companies. These companies are Chevron, McDonalds, Microsoft, NIKE, salesforce, Tesla, and Walmart. Across all companies, it is found that the significant effects of Twitter-based variables on conditional volatility mostly occur during economic downturn, which is consistent with Hsu et al. (2021). However, for model II, the Twitter-based sentiment and the standardised number of interactions are significant more often for $G(\cdot) = 0$. Under $G(\cdot) = 1$, the standardised number of interactions is significant more often. In the case of Apple, Twitter-based variables explain the volatility of the stock price returns better in economic prosperity, which contrast the intuition.

Unlike the constant parameter case, the varying parameter version of model III does not provide substantial benefits over the varying parameter version of model II. Especially during economic prosperity, the effect of the Twitter variables on the volatility process are found to be significant across less companies. Although the varying parameter specification provides a more flexible structure, and this increases how often Twitter variables can successfully explain the volatility process, this model does not incorporate Twitter-based variables more successfully than the varying parameter version of model II. More specifically, based on the parameter estimates, it is argued that only for Coca-Cola, McDonalds, Microsoft, NIKE, and Walmart, the varying parameter specification of model III provides the best model to include exogenous variables. Based on the plots of appendix A.2.1, this is in contrast with the expectation, especially in the case of Microsoft and NIKE.

Across both varying parameter models, negative sentiment has a larger effect on the conditional volatility than positive sentiment, which supports the expectation. Additionally, in contrast to the findings by Audrino et al. (2020) and Rakowski et al. (2021), based on the parameter estimates, there is no apparent relationship between the distribution of stock ownership and the effects of Twitter-based sentiment.

6.2 Model comparison

In order to determine which conditional volatility model specification incorporates the exogenous Twitter-based variables via the best methodology, several tests were proposed in section 4.5. These test are conducted to determine the *best-practice* approach when incorporating Twitter-based variables into the conditional volatility equation of stock price returns. Moreover, it is investigated whether allowing for a varying parameter specification based on the state-of-the-economy can significantly benefit the in-sample fit of the proposed volatility models. Finally, the in-sample fit of these models is evaluated using a ranking algorithm based on the MSE and QLIKE loss scores, which are computed using a volatility proxy $\hat{\sigma}^2_{t,prox}$.

To provide some intuition for the behaviour of the conditional volatility processes, figure 2 shows the conditional volatility of the proposed models for Apple, Amazon, Chevron and Coca-Cola. Figure 6 and figure 7 in appendix A.2.2 show the plots of the conditional volatility processes for the remaining companies in this research. These plots show for each company the conditional volatility process of each model, the varying-parameter counterpart and the benchmark volatility process specified by the parameters from table 14 in appendix A.1.3. Recall that the squared residuals ε_t^2 of the mean returns series are used as a proxy for the conditional variance $\hat{\sigma}^2_{t,prox}$.

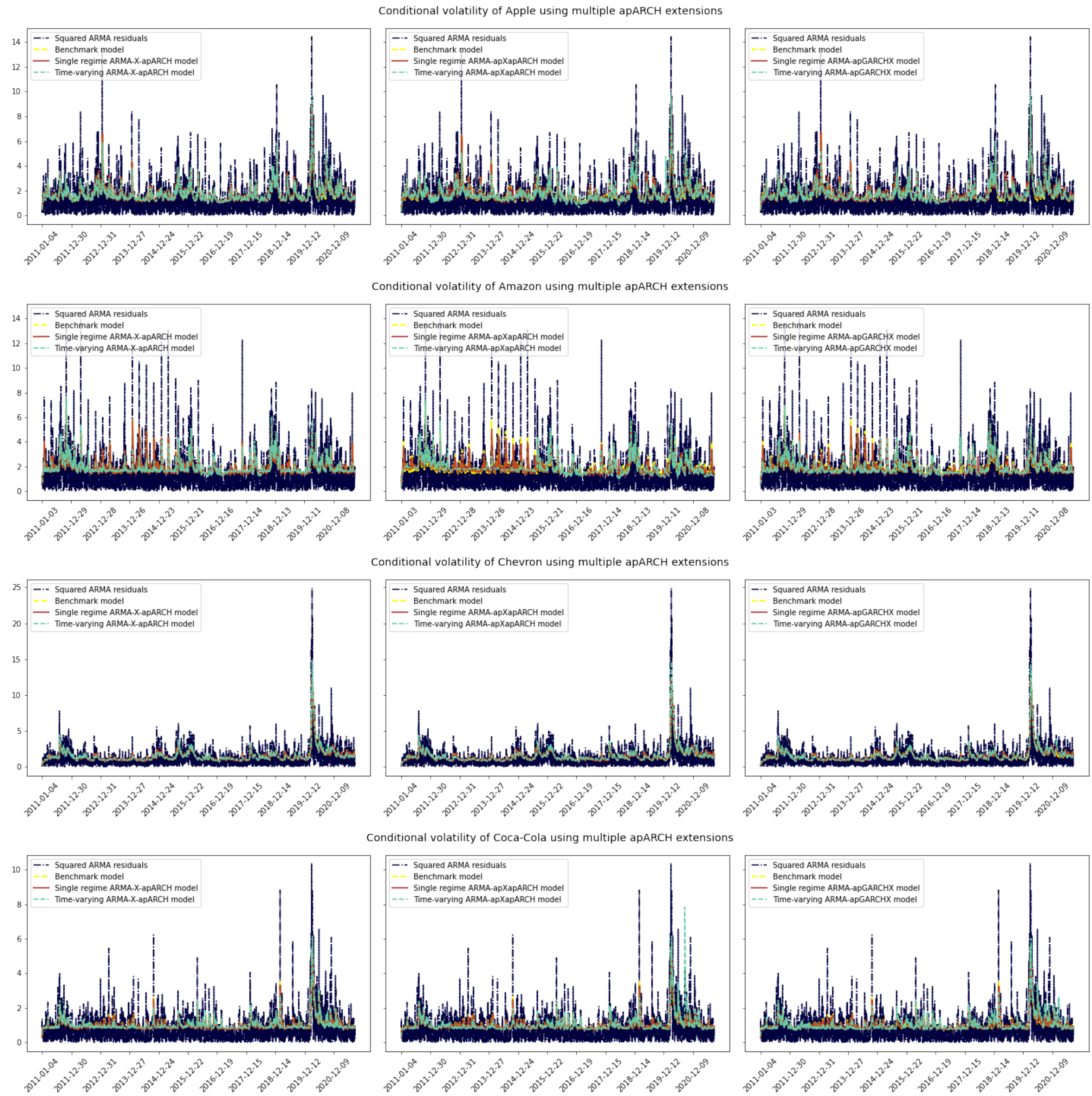
As described in section 4.5, it is mainly considered how well the conditional volatility processes described $\tilde{\sigma}_t$ captures the spikes in the volatility, since failure to capture these high values of volatility can induce excessive risk taking. From the plots, there is no readily derived model that significantly captures the conditional volatility better for all companies. Recall from table 3 that the standard deviation of the stock price returns series for Netflix and Tesla was much higher compared to the other companies. This observation indicates high levels of volatility throughout the series, this is confirmed by the plots. Yet, the high standard deviation of Netflix is caused by a few outliers, whereas the volatility of Tesla is generally high and has a lot of spikes throughout the entire sample.

Consistent with the parameter estimates, the plots show that the ability of a model to capture volatility spikes is company specific. The time-series plots show favorability for the constant parameter specification for Amazon, Coca-Cola (to lesser extent), and Microsoft, since it can be observed that the constant parameter version are better to capture volatility spikes. This contradicts the results from the parameter estimates, which show that in the varying parameter setting, there are more Twitter-based variables that successfully explain the conditional volatility series for these companies.

Additionally, consistent with the parameter estimates, the plots reveal that indeed for Coca-Cola and Walmart, model III seems to provide a better fit than model II. This evidence is less convincing for McDonalds, Microsoft, and NIKE. Subsequently, consider the plots from appendix A.2.1, which display the volatility of the exogenous variables. From these plots and figure 1, it can be seen that for Coca-Cola, the spikes in the volatility of the attention variables in 2015 and 2019 can be used to explain the volatility of Coca-Cola during economic downturn. Similarly, it can be concluded that spikes in the volatility of the attention variables of Walmart occur when $G(\cdot) = 1$, and are therefore useful to explain increased conditional volatility. These relationships are less evident for the other variables.

For the other companies, the plots show that model II provides the best fit. In general, the varying parameter models seem to better capture volatility clustering and model volatility spikes. However, by simultaneously considering figure 1, it can be deduced that periods of increased volatility of the stock price returns do not necessarily correlate with the VIX and the TED rate. This only holds for the period around the outbreak of COVID-19, and for the volatility of Chevron, Exxon Mobil, McDonalds, salesforce, and Walmart (to lesser extent). This implies that the period where the structural breaks occur based on the function $G(\cdot)$ is inconsistent with volatility clusters.

Figure 2 – Conditional volatility between 01-01-2011 and 31-08-2021 of U.S. companies stock price returns.



Time series plot of the conditional volatility for Apple, Amazon, Chevron and Coca-Cola (top to bottom plots) between 01-01-2011 and 31-08-2021. Squared ARMA residuals serves as a proxy for the conditional variance process (square root of proxy is denoted in dark blue). From left to right, plots show time series plots of the conditional volatility using the ARMAX-apARCH model, the ARMA-apARCH-apX model, and the ARMA-apARCH-apXGARCH model, for both the constant and time-varying parameter specification. In yellow, the conditional volatility of the benchmark model is plotted.

6.2.1 Quasi Likelihood Ratio test

The previous subsections discussed the time series plots of the conditional volatility process and the parameter estimates for the models that include Twitter-based variables via the specifications in section 4.1. To test whether the inclusion of Twitter-based exogenous variables significantly increases the quasi log-likelihood, the Quasi Likelihood Ratio test, given by equation (36) is performed. Later, the results from this test are compared with the minimal value of the BIC, to see whether the selected model is consistent based on these evaluation metrics. Table 10 shows for each company included in this paper the quasi log-likelihood $\mathcal{L}_T$ under the parameters that minimize equation (28), and the BIC based on this value $\mathcal{L}_T$. Moreover, the QLR statistics Λ^i_{QLR} under various null hypotheses i are given.

Recall that for large values of the test statistic Λ^i_{QLR} the null is rejected, indicating that the unrestricted model significantly increases the quasi log-likelihood compared to the restricted model. The QLR test statistics are iteratively considered for $i = 1, 2, 3$ to determine which model provides the best fit.

For $i = 1$, the null hypothesis states that the parameters which include the Twitter-based exogenous variables are zero. This is equivalent to testing whether the quasi log-likelihood $\mathcal{L}_T$ of any model that includes exogenous variables is significantly larger than the quasi log-likelihood from the benchmark model. Using equation (36), the quasi log-likelihood of the benchmark model is compared against the quasi log-likelihood of the proposed models, where the degrees of freedom correspond to the number of elements that is assumed zero under the null.

Table 10 shows that Λ^1_{QLR} is significant at at least the 10% level for the constant parameter models of Apple, Amazon, Chevron, and NIKE. This indicates that for these companies, the null is successfully rejected at the 10% significance level, which indicates that the inclusion of exogenous variables significantly increases the quasi log-likelihood for all models. For McDonalds, Netflix, and salesforce, under the conditional volatility specification from equation (23), the null hypothesis is significantly rejected. Moreover, for Microsoft, Netflix, salesforce, and Tesla, the conditional volatility under equation (20) significantly increases the quasi log-likelihood. Additionally, the null is rejected for all companies, except Tesla, which proves that the varying parameter models perform significantly better than the benchmark model. Still, this does not imply that the inclusion of the exogenous covariates in a varying parameter setting significantly increases the quasi log-likelihood.

Subsequently, it is tested whether the models that allow for varying parameters outperform the models that assume constant parameters. To this extent, the test statistic Λ^2_{QLR} is computed. This test statistic is calculated via equation (36) under the null hypothesis that all elements in θ^* are equal to zero. That is, it is tested whether allowing for varying parameters in the conditional volatility equation can significantly increase the quasi log-likelihood compared to the constant parameter counterparts. The null is rejected at at least the 10% significance level for all models, except for Tesla. This implies that accommodating for varying parameters in the conditional volatility equation significantly improves the models that assume constant parameters. This is consistent with the parameter estimates from table 7, table 8, and table 9 and the tables with the estimated varying parameters in appendix A.1.4. It was established that under the specification of section 4.4, the number of significant parameter estimates that include the exogenous variables increased substantially.

Finally, the test statistic Λ^3_{QLR} compares both the constant and varying parameter specifications of model II against model III. This is equivalent to testing H_0: $\lambda_{k,i} = 0$ and $\lambda^*_{k,i} = 0$ for all $i \in \{1, \dots, S_k\}$, $k \in \{1, \dots n_{K_1}\}$.

This test is conducted to infer whether the conditional volatility under equation (23) significantly improves the fit compared to the volatility under equation (20). From table 10, it is deduced that for Coca-Cola, McDonalds, NIKE, and Walmart, the null is rejected at at least the 10% level, both for the constant as well for the varying parameter specification. This finding is consistent with the aforementioned conclusion regarding the parameter estimates.

Table 10 – Quasi maximum likelihood, BIC and LR statistics of the conditional volatility processes (Table continues on next page).

	Constant parameter				***Varying parameter specification***		
	Benchmark	Model I	Model II	Model III	Model I	Model II	Model III
AAPL							
$\mathcal{L}_T$	-5.253e+03	-5.241e+03	-5.226e+03	-5.231e+03	-5.237e+03	-5.204e+03	-5.212e+03
BIC	2.819e+07	2.812e+07	2.804e+07	2.807e+07	2.810e+07	2.792e+07	2.797e+07
Λ^1_{QLR}		24.134***	54.314***	44.33***	32.757***	98.84***	81.591***
Λ^2_{QLR}					8.624*	44.527***	37.26***
Λ^3_{QLR}			-9.983			-17.25	
AMZN							
$\mathcal{L}_T$	-6.020e+03	-6.011e+03	-5.985e+03	-6.000e+03	-5.977e+03	-5.921e+03	-5.961e+03
BIC	3.232e+07	3.227e+07	3.213e+07	3.221e+07	3.208e+07	3.179e+07	3.200e+07
Λ^1_{QLR}		17.267***	70.866***	40.083***	86.362***	197.604***	117.82***
Λ^2_{QLR}					69.095***	126.738***	77.737***
Λ^3_{QLR}			-30.782			-79.784	
CVX							
$\mathcal{L}_T$	-4.331e+03	-4.323e+03	-4.323e+03	-4.319e+03	-4.310e+03	-4.285e+03	-4.288e+03
BIC	2.324e+07	2.320e+07	2.320e+07	2.318e+07	2.313e+07	2.299e+07	2.301e+07
Λ^1_{QLR}		16.798***	15.475**	24.057***	42.933***	91.889***	85.701***
Λ^2_{QLR}					26.136***	76.414***	61.644***
Λ^3_{QLR}			8.582**			-6.188	
KO							
$\mathcal{L}_T$	-2.583e+03	-2.581e+03	-2.581e+03	-2.578e+03	-2.542e+03	-2.512e+03	-2.492e+03
BIC	1.386e+07	1.385e+07	1.385e+07	1.383e+07	1.364e+07	1.348e+07	1.337e+07
Λ^1_{QLR}		3.798	3.454	9.876	81.72***	141.107***	181.632***
Λ^2_{QLR}					77.922***	137.653***	171.755***
Λ^3_{QLR}			6.423*			40.525***	
XOM							
$\mathcal{L}_T$	-3.875e+03	-3.948e+03	-3.870e+03	-3.869e+03	-3.865e+03	-3.851e+03	-3.859e+03
BIC	2.080e+07	2.118e+07	2.077e+07	2.076e+07	2.074e+07	2.066e+07	2.071e+07
Λ^1_{QLR}		-144.35	10.028	11.932	20.92***	48.902***	33.511*
Λ^2_{QLR}					165.27***	38.873***	21.579*
Λ^3_{QLR}			1.904			-15.391	
MCD							
$\mathcal{L}_T$	-2.719e+03	-2.721e+03	-2.719e+03	-2.712e+03	-2.663e+03	-2.662e+03	-2.654e+03
BIC	1.459e+07	1.460e+07	1.459e+07	1.455e+07	1.429e+07	1.428e+07	1.424e+07
Λ^1_{QLR}		-3.785	1.036	14.978*	111.725***	114.555***	130.132***
Λ^2_{QLR}					115.51***	113.519***	115.154***
Λ^3_{QLR}			13.942***			15.577**	
MSFT							
$\mathcal{L}_T$	-4.644e+03	-4.619e+03	-4.630e+03	-4.642e+03	-4.596e+03	-4.568e+03	-4.583e+03
BIC	2.492e+07	2.478e+07	2.484e+07	2.491e+07	2.466e+07	2.451e+07	2.459e+07
Λ^1_{QLR}		51.806***	29.2***	4.853	97.077***	153.678***	123.551***
Λ^2_{QLR}					45.271***	124.478***	118.698***
Λ^3_{QLR}			-24.347			-30.127	

	Constant parameter				Varying parameter specification		
	Benchmark	Model I	Model II	Model III	Model I	Model II	Model III
NFLX							
$\mathcal{L}_T$	-8.629e+03	-8.987e+03	-8.509e+03	-8.617e+03	-8.489e+03	-8.441e+03	-8.435e+03
BIC	4.630e+07	4.822e+07	4.566e+07	4.624e+07	4.555e+07	4.529e+07	4.526e+07
Λ^1_{QLR}		-715.915	239.766***	23.447***	281.118***	376.452***	388.075***
Λ^2_{QLR}					997.033***	136.686***	364.628***
Λ^3_{QLR}			-216.319			11.623*	
NKE							
$\mathcal{L}_T$	-4.918e+03	-4.912e+03	-4.912e+03	-4.878e+03	-4.881e+03	-4.832e+03	-4.797e+03
BIC	2.639e+07	2.636e+07	2.636e+07	2.617e+07	2.619e+07	2.593e+07	2.574e+07
Λ^1_{QLR}		10.991**	11.43*	79.932***	73.427***	171.843***	242.463***
Λ^2_{QLR}					62.437***	160.413***	162.531***
Λ^3_{QLR}			68.502***			70.62***	
CRM							
$\mathcal{L}_T$	-6.407e+03	-6.479e+03	-6.279e+03	-6.394e+03	-6.432e+03	-6.231e+03	-6.303e+03
BIC	3.438e+07	3.477e+07	3.369e+07	3.431e+07	3.452e+07	3.344e+07	3.382e+07
Λ^1_{QLR}		-144.992	255.18***	24.869***	-51.087	350.614***	208.212***
Λ^2_{QLR}					93.905***	95.434***	183.343***
Λ^3_{QLR}			-230.311			-142.402	
TSLA							
$\mathcal{L}_T$	-9.031e+03	-9.044e+03	-9.011e+03	-9.030e+03	-9.029e+03	-9.014e+03	-9.018e+03
BIC	4.848e+07	4.855e+07	4.837e+07	4.847e+07	4.847e+07	4.839e+07	4.841e+07
Λ^1_{QLR}		-26.61	38.668***	1.161	2.758	32.998**	25.575
Λ^2_{QLR}					29.368***	-5.67	24.414**
Λ^3_{QLR}			-37.507			-7.423	
WMT							
$\mathcal{L}_T$	-3.266e+03	-3.261e+03	-3.299e+03	-3.263e+03	-3.144e+03	-3.210e+03	-3.126e+03
BIC	1.753e+07	1.751e+07	1.771e+07	1.751e+07	1.688e+07	1.723e+07	1.678e+07
Λ^1_{QLR}		9.048**	-67.352	6.158	243.912***	111.464***	278.604***
Λ^2_{QLR}					234.864***	178.816***	272.447***
Λ^3_{QLR}			73.51***			167.14***	

Table presents for each company the maximized Quasi Log Likelihood and the BIC.
[1]: Λ^1_{QLR} is the test statistic under the null that the inclusion of exogenous variables does not improve the model.
[2]: Λ^2_{QLR} is the test statistic under the null hypothesis that the varying elements are equal to zero, *i.e.*, H_0: $\theta^* = \mathbf{0}_{m_2}$ where m_2 denotes the number of elements in θ^*.
[3]: Λ^3_{QLR} is the test statistic under the null that the *disagreement measures* $\sigma^{(x)}_t$ have no effect on the conditional volatility.
Asterisks (*, **, ***) denote significance of the QLR statistic at the 10%, 5% and 1% significance level, respectively, calculated using the χ^2 distribution.

By simultaneously considering the test statistics in table 10, the following can be concluded. For Coca-Cola, McDonalds, NIKE, and Walmart, the varying-parameter specification of model III is considered the most useful to explain conditional volatility using Twitter-based variables. For Tesla, the volatility specified by the ARMA-apARCH-apX model from equation (20) provides the highest quasi log-likelihood. For all the other companies, the model that yields the best likelihood that is significantly higher than any other specification is the varying-parameter specification of model II. These findings are consistent with the BIC reported in the table. Recall that the BIC provides a weigh-off between model complexity and the likelihood. This supports the finding that the Quasi Likelihood Ratio tests correctly identified the best model for each company

6.2.2 Evaluation of the in-sample fit

Lastly, the in-sample fit of the volatility models are compared via the method described in section 4.5.3. Due to the fact that the conditional volatility is an unobservable process, the estimated processes are compared to a volatility proxy. As aforementioned, since this proxy is not the 'true' volatility, evaluation via the loss functions provides a qualitative assessment of the in-sample fit, rather than a quantitative assessment. This implies that no remarks are made on the relative size of the errors computed by the loss functions.

Recall from section 4.5.3 that Patton (2011) identified two robust evaluation metrics, in the sense that the ranking of multiple models by these loss functions is consistent regardless of the volatility proxy $\hat{\sigma}^2_{t,prox} = \tilde{\varepsilon}^2_t$. As aforementioned, the MSE is a symmetric function, that equally values negative and positive losses of the same magnitude, and the QLIKE is asymmetric and 'punishes' positive losses less severe than negative losses of the same magnitude.

From the conditional volatility plots in figure 2, and figure 6 and figure 7 in appendix A.2.2 it can be seen that across all companies, the level of the conditional volatility tends to be much higher than the level of the volatility proxy, which causes positive errors. During periods with increased levels of volatility, the conditional volatility is generally lower than the proxy, which causes negative loss. The errors, computed by the MSE and the QLIKE loss functions are tabulated in table 11. The numbers in bold denote the lowest error, and indicate which model provides the best in-sample fit for each company.

Table 11 – In-sample loss scores for different models. Data ranges from 01-01-2011 until 31-08-2021

	Loss scores of different models against volatility proxy											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
	Mean Squared Error											
Benchmark	81.84	131.85	263.78	15.19	55.42	67.76	64.07	2845.12	96.34	**252.99**	1068.29	35.54
Model I	81.64	132.05	263.50	15.17	55.37	**67.01**	64.30	3033.83	96.28	259.57	1070.36	35.52
Model II	82.96	130.56	263.53	15.16	55.44	67.79	64.07	**2808.60**	96.44	259.77	1068.91	35.91
Model III	82.17	131.84	**263.06**	15.15	55.43	68.20	64.20	2827.21	96.54	305.28	1067.78	35.49
vp Model I	81.54	131.54	265.16	14.78	55.41	72.80	**63.86**	2967.69	**95.49**	255.04	1073.75	35.84
vp Model II	82.51	**130.28**	267.85	18.16	55.36	72.29	63.87	2839.85	95.90	260.91	**1061.35**	**34.86**
vp Model III	**81.33**	131.26	264.44	**15.07**	**55.28**	72.84	64.34	2942.07	95.68	522.80	1073.76	35.40
	QLIKE loss											
Benchmark	1.958	2.243	1.614	0.963	1.444	1.014	1.731	3.216	1.833	2.388	3.365	1.217
Model I	1.955	2.240	1.614	0.962	1.473	1.015	1.727	3.232	1.833	2.413	3.367	1.217
Model II	1.948	2.230	1.612	0.962	1.443	1.013	1.726	3.192	1.831	2.343	3.357	1.231
Model III	1.950	2.236	1.611	0.961	1.442	1.011	1.730	3.226	1.818	2.385	3.364	1.216
vp Model I	1.954	2.228	1.598	0.948	1.443	0.993	1.719	3.156	1.822	2.393	3.360	1.174
vp Model II	**1.939**	**2.206**	**1.598**	0.936	**1.435**	0.992	**1.703**	3.177	1.802	**2.325**	**3.354**	1.199
vp Model III	1.943	2.221	1.599	**0.929**	1.438	**0.989**	1.708	**3.141**	**1.788**	2.350	3.360	**1.166**

Errors computed by the MSE and QLIKE loss functions for all models per company. Loss scores are calculated using a variance proxy $\hat{\sigma}^2_{t,prox}$ to compare with the estimated volatility by the models. Volatility proxy is the square root of the squared residuals from the benchmark model. Bold cases denote the best score per company for the MSE and QLIKE errors.

Since the QLIKE loss function is asymmetric, the ability of a conditional volatility model to correctly model volatility spikes is more important than the ability to model low values of volatility. While models that are unable to properly account for high values of volatility can indulge excessive risk taking by the user, the risks for overestimating the actual volatility bears less severe risks. The QLIKE 'punishes' the losses accordingly and values the ability to capture spikes in the volatility more. Hence, the error computed by the QLIKE loss function serves as the most important loss function.

From the values of the MSE, it is clear that no model unanimously performs better than any other model across the selected companies. Nonetheless, it must be noted that the errors for the models that allow for varying parameters are generally smaller than the errors of the constant parameter models. This is consistent with the conclusions from the Quasi Likelihood Ratio test, and supported by the time series plots of the conditional volatility. By inspecting the conditional volatility plots, there is no apparent reason why the MSE is lower for the constant parameter specification of McDonalds, Neftflix, NIKE and salesforce.

Subsequently, the errors computed by the Quasi Likelihood loss functions are the smallest for the models that allow for varying parameters. As aforementioned, this loss functions punishes negative errors more severe than positive errors, which is beneficial when testing the in-sample fit of volatility models since underestimating the conditional volatility can lead to excessive risk taking, and overestimating the volatility will merely decrease risk appetite. From the table, it can be deduced that across all companies the models that allow for varying parameters are ranked better than the models that assume constant parameters.

It is remarkable that the ranking algorithm using the QLIKE loss function prefers the same models for almost all companies as the QLR test statistic. Only for Netflix and Tesla the ranking is different, which could be due to the fact that these companies may experience more stock price return volatility than the other companies in the sample, which is supported by the descriptive statistics in table 3.

Hence, by evaluating the in-sample fit of the models using the QLIKE and MSE error, it can be concluded that the results are consistent with the parameter estimates and the results of the Quasi Likelihood Ratio test. From these results, it can be concluded that in general, models that allow for varying parameters in the volatility equation provide a better fit than their constant parameter counterparts. Additionally, in line with the literature, Twitter-variables successfully explain the conditional volatility better in times of economic downturn. From the parameter estimates, it is concluded that, consistent with Smales (2014), negative Twitter-sentiment has a larger effect on the volatility series, and this asymmetry increases during economic downturn. In contrast with previous findings, the significant effects of the Twitter-based sentiment and the number of interactions decreases under economic distress, whereas the significant effects of the standardised number of tweets increases when the VIX and TED rate are high.

Moreover, it can be concluded that for customer intensive[15] companies, the volatility of the Twitter-based variables exerts significant effects on the volatility. This holds for Coca-Cola, McDonalds, NIKE, and Walmart. For the other companies, (the varying parameter specification of) model II is proven to provide the best in-sample fit, and significantly improves the fit that does not include exogenous variables.

Lastly, in contrast to the findings of Audrino et al. (2020) and Rakowski et al. (2021), there is no evidence that Twitter-based sentiment affects companies with a large portion of retail investors more than companies with little percentage of retail investors.

[15]This entails companies that belong to the S&P 500 Consumer Discretionary Index or the S&P500 Consumer Staples Index.

7 Summary

This chapter discusses the methodologies used throughout this study, the main findings are discussed, and their implications are explained in more detail. This research aims to determine how Twitter activity can be leveraged to explain stock price volatility. Additionally, it is investigated how long these effects of Twitter persist, and whether these effects are different during different states-of-the-economy. Since it is argued that these stock price fluctuations are mostly attributed to the behaviour of retail investors, this paper specifically considers the effect of non-financial tweets.

To this extent, this paper uses the Twitter API to retrieve tweets of 12 major U.S. companies. The content of these tweets is specifically selected to address non-financial issues, such as bad working conditions and environmental concerns. Moreover, this study specifically considers the hazards of spam tweets, and aims to limit the amount of spam tweets by filtering out words that do not display emotions towards the selected companies. In contrast to previous research (Fan et al., 2020; Ramco et al., 2015), this research develops a method to aggregate multiple tweets within a certain period. This method is based on the article by Perdana and Pinandito (2018), who employ Twitter-specific public metrics to derive a non-textual importance for each tweet. This research argues that the number of retweets, quotes, likes, and replies of a tweet imply the relative influence a tweet generates. To this extent, the method of Perdana and Pinandito (2018) is modified to assign each tweet within an interval a non-textual weight. This weight is subsequently used to derive a weighted average sentiment score, based on multiple tweets send between an interval. This approach simultaneously deals with tweets that are not credible and not worthwhile considering, since these tweets are assigned a lower weight when they do not generate any interactions. Hence, this method also deals with filtering out spam tweets, without the consideration of extra data.

To determine the sentiment that the selected tweets display, the sentiment polarity of the tweets is calculated. Unlike previous research, this paper performs a study to determine an appropriate sentiment classification method. Since the performance of sentiment classification methods is very context dependent, this research tests several methods on a subsample of the retrieved Twitter data. To construct this test dataset, a panel of eight peer students manually annotated 2200 tweets. The construction of a ground-truth test dataset, with context specific tweets, allows to test multiple classification methods. From the performance of the methods on this manually annotated subsample, it can be inferred how well each method is likely to perform on the entire Twitter dataset.

This study aims to employ either the VADER lexicon based approach, constructed by Hutto and Gilbert (2015), or the Naive Bayes machine learning approach. The NB performs generally well on classification assignments (Go et al., 2009), and is trained on a training set constructed using multiple publicly available datasets. Besides the VADER lexicon-based approach, this thesis proposes to extend this lexicon by adding domain specific words to this lexicon. It is proposed that adding domain specific words (*e.g.*, climate change, wage gap) increases the classification accuracy of the VADER lexicon on the retrieved tweets. As proposed by Yang (1999), the Macro F_1 score and the accuracy of these sentiment classification methods are computed to evaluate the performance of these methods on the manually annotated test set. It is concluded that the adjusted VADER makes the most precise classifications, although a substantial fraction of tweets in the test set is still missclassified. However, this extensive study does provide insights in how readily available classification methods generally perform.

Moreover, this study also accounts for possible effects of the number of tweets send per day, and the number of interactions per day, in line with the suggestions by Audrino et al. (2020).

Subsequently, the appropriate modeling technique is chosen that allows for the inclusion of these variables. In financial time series, it is common that volatility tends to cluster and depends on past levels of the volatility. These characteristics are captured by the GARCH model developed by Bollerslev (1986). Over the years, multiple alternatives to these types of models have been proposed, of which several allow for the inclusion of exogenous variables. Since it is argued that the mean stock price return does not necessarily follow a pure GARCH model (Franq & Zakoian, 2004), an auto-regressive moving-average model describes the mean return series, where the innovations follow an apARCH model. This model, developed by Ding et al. (1993), provides an extension of the GARCH model that takes into account the leverage effect observed by Black (1976). This effect describes the tendency that negative shocks have larger effects on the volatility than positive shocks.

Using the apARCH model as a benchmark, this paper introduces three main extensions, that allow for Twitter-variables to explain the volatility process. Ultimately, it is explored whether accounting for Twitter variables can improve the in-sample fit compared to the benchmark case. Unlike other papers, this thesis focuses on the in-sample fit, and no inferences are made about the forecasting power of the suggested models. However, this research does provide meaningful suggestion to include Twitter to explain volatility better, which could be easily extended to make Out-of-Sample forecasts.

The proposed models in this research introduce the Twitter-based variables via different methodologies. In the first place, the Twitter-based variables enter the mean return equation. Next, two models are introduced were the Twitter-based variables enter the conditional volatility equation. Similarly to the leverage effect, these variables enter the models asymmetrically, to infer whether negative sentiment has a different effect on the volatility process than positive sentiment. Lastly, the third model additionally includes the conditional volatility of these exogenous variables as extra explanatory variable in the conditional volatility equation of the stock price returns. These measures serve as disagreement measures and describe how much the Twitter-based variables vary over time.

Additionally, a robustness check is proposed to determine whether the effects of Twitter are different through different states of the economy, which is supported by Hsu et al. (2021). To this extent, a smooth transition function is introduced, inspired by the function by Amado and Teräsvirta (2008), that assumes value zero or one, based on the state of the economy. To model the state of the economy, this thesis uses the VIX and the TED rate. For high values of these variables, the smooth transition function takes value one, and zero elsewhere. Hence, it can be investigated whether Twitter variables can better explain conditional volatility when their effects are dependent on the state of the economy.

Finally, the volatility processes are estimated using the Quasi Maximum Likelihood Estimator. Franq and Thieu (2019) and Franq and Zakoian (2004) provide the required assumptions that are necessary for consistent estimation of the 'true' model parameters. The in-sample comparison of the different models is done by means of the Quasi Likelihood Ratio tests, and the in-sample fit is evaluated using the QLIKE and MSE loss functions.

In this research, it is found that the Twitter-based variables can significantly improve traditional models. Based on the Bayes Information Criterion, it is deduced that in general, Twitter-based variables only effect the one-day ahead stock price return volatility, which is consistent with Audrino et al. (2020). Based on the significance of the parameter estimates, the results from the QLR tests, and the evaluation of the in-sample fit, it is shown that Twitter variables successfully help to explain the volatility series. Consistent with the findings of Audrino et al. (2020), it is found that Twitter-based sentiment asymmetrically affects the conditional volatility process. More specifically, it is found that negative sentiment has more effect on the conditional volatility than positive sentiment. The magnitude of this asymmetry increases during economic downturn. Additionally, by using the varying parameter specification, it is found that during economic downturn, the effects of Twitter on conditional volatility are more significant, which is consistent with Hsu et al. (2021). Hence, accounting for varying parameters and asymmetric effects of the Twitter variables can significantly improve the in-sample fit of conditional volatility models.

In the varying parameter setting, inclusion of the volatility of the exogenous variables is shown to result in better volatility models for a limited number of companies. This only holds for companies that are very customer intensive. However, in general it might provide valuable to include a certain *disagreement measure*, similar to Fan et al. (2020). In the constant parameter setting, the volatility is better explained by models that include both the Twitter variables as well as the volatility of these variables.

8 Conclusions and recommendations

This study considers the effect on non-financial Twitter activity on stock price return volatility. It is explored how Twitter sentiment, and Twitter-based attention measures influence stock price returns volatility. This study explores multiple approaches to meaningful derive the methods via which Twitter affects the volatility of stock price returns. The Twitter data consists of non-financial related tweets covering 12 major U.S. companies between 01-01-2011 and 31-08-2021. During this period, social media is gradually adopted in our daily lives, moreover, this period covers multiple recessions. Additionally, this research considers multiple sentiment classification methods and introduces a parsimonious method to derive a daily weighted average Twitter sentiment score.

To account for the effects of Twitter, this research carefully considers multiple sentiment classification methods. In line with previous research (Ribeiro et al., 2016), it is found that the classification methods are very domain sensitive, and generally fail to succesfully classify a large part of the retrieved tweets. To overcome this obstacle, this research aims to improve the VADER sentiment lexicon, by taking into account the specific domain of the retrieved tweets. However, the classification is still relatively poor. Hence, it is concluded that readily available sentiment lexicons do not perform very well in classifying very domain specific data. This implies that careful consideration must be paid to the classification of the tweets. This could also imply that the role of the public sentiment, calculated using non-financial tweets, is underestimated in this research. Therefore, it is suggested that in future research, more resources must be consulted to construct an appropriate training dataset, as machine-learning models generally perform better in classifying sentiment (Hartmann et al., 2019). If this is impossible, the researchers should consider tuning readily available lexicons to the domain on a larger scale, using methods similar to those proposed in this research.

Moreover, this paper presents a parsimonious method to distinguish important and non-important tweets, based on public metrics associated with each tweet. Unlike other research, this approach does not require additional information other than the information already stored with each tweet. This method offers a substantial improvement compared to earlier research, where the sentiment scores of tweets were simply averaged to derive the daily sentiment score.

However, it is not tested whether this weighted average method actually succeeds to derive a superior proxy for the sentiment score, which in turn has more significant effect on the conditional volatility. Future research could investigate whether there are observable differences using this aggregation algorithm, in contrast to simple averaging of polarity scores.

After these variables are extensively reviewed, it is tested whether Twitter-based variables can successfully explain the volatility of stock price returns. In conclusion, it is found that under specific circumstances, Twitter exerts significant influence on stock price return volatility. This is shown by multiple quasi log-likelihood tests, as well as by the error computed by multiple loss functions, which are calculated on the in-sample fit of the conditional volatility models. Consistent with the findings of Audrino et al. (2020) and Smales (2015), it is found that negative Twitter based sentiment has a larger effect on the conditional volatility than positive Twitter based sentiment. Additionally, as proposed by Audrino et al. (2020), this study explores whether the effects of Twitter are different over time. To this extent, parameters are allowed to vary as a function of macroeconomic and market-based indicators. This function is inspired by Amado and Teräsvirta (2008). The existence of varying parameters are confirmed, and in line with Hsu et al. (2021), it is concluded that under worsening economic conditions, Twitter-based variables have a larger effect on the conditional volatility. In contrast to Audrino et al. (2020) and Rakowski et al. (2021), there is no evidence that Twitter-based variables particularly affect stocks that are primarily traded by retail investors. When the volatility of the Twitter-based variables enter the conditional volatility equations, inspired by Fan et al. (2020), it is concluded that in the constant parameter setting, the significance of these exogenous variables substantially increases across all companies included in this research. In the varying parameter setting, only for Coca-Cola, McDonalds, NIKE, and Walmart, the volatility of the Twitter-based variables show significant effect on the conditional volatility process. In contrast with Kim and Kon (1999) and Lamoureux and Lastrapes (1990) it is concluded that during economic downturn, the conditional volatility process exhibits a higher degree of persistence compared to the constant parameter case. This indicates that when Twitter-based variables are used to explain the volatility of stock price returns, shocks to this process take longer to decay over time.

However, this research fails to test the proposed methodologies to make meaningful volatility forecasts, which is the goal of volatility models in almost all financial applications (Patton, 2011). Thus, future research could investigate whether Twitter-based variables also have predictive power, and can significantly explain future volatility.

Finally, this research only considers lagged values of the Twitter-based variables to affect the volatility process. Yet, Hsu et al. (2021) show that accounting for the contemporaneous effects of news can also successfully explain volatility, and improve volatility forecasting. Hence, future research should consider contemporaneous effects of Twitter sentiment as well as lagged effects, especially considering the fact that information dispersion through social media is rather instantaneous compared to more traditional news sources.

All in all, this research shows several practices that can be employed to successfully model volatility using Twitter-based variables. Volatility models that use these variables can capture Twitter attention and the public attitude towards certain companies. Given the increase in retail investor activity, illustrated by the GameStop Mania (Fitzgerald, 2021), professional investors and pension funds can account for the effects of retail investor behaviour by employing the proposed methodologies. Additionally, this research can be used by companies to steer away from controversies, as it is argued that negative sentiment can significantly increase volatility, which makes these companies less attractive for professional investors.

A Appendix

A.1 Tables

A.1.1 Twitter API

Table 12 – ESG dictionary used for the Twitter API

ESG dictionary	eco-friendly, harassment, fraud, sustainable, governance, responsible, beneficial, unhealthy, pollution, climate change, clean energy, carbon, ethical, green energy, green ambitions, greenhouse gas, exclusions, negative, human-rights, human rights, impact investing, slavery, renewable, energy transition, sustainable transition, SDG, unfair, discrimination, sexism, ESG, CSR, green development, inclusion, waste, favoritism, impoverished, obesity, scarcity, layoff, unemployment, renewables, inclusion, employee bonus

List of words that are included in the *search query*. Table shows both the constructed *ESG dictionary*, containing words associated with the non-financial performance of corporations. All tweets in the retrieved data must contain at least one of the words in this dictionary.

A.1.2 List of market capitalization and % of shares held by institutional investors

Table 13 – Company characteristics of selected U.S. companies

Company name	Market capitalization	Percentage of shares held by institutional investors	Percentage of shares held by insiders
Apple	2.46T	58.88%	0.07%
Amazon	1.668T	59.15%	13.62%
Chevron	197.18B	68.42%	0.03%
Coca-Cola	247.04B	68.39%	0.64%
Exxon Mobil	240.34B	52.85%	0.12%
McDonalds	178.35B	69.08%	0.05%
Microsoft	2.2T	72.07%	0.08%
Netflix	228.34B	82.39%	1.53%
NIKE	271.8B	83.44%	1.22%
salesforce	244.4B	79.22%	3.62%
Tesla	710.01B	41.89%	18.96%
Walmart	419.005B	30.92%	49.28%

List of market cap and percentage of shares held by institutional investors at 02-08-2021 for companies in this research. Data collected via the yahoo! finance website

A.1.3 Benchmark model parameter estimates

This section presents the parameter estimates of the benchmark model, that does not include exogenous covariates.

Table 14 – Parameter estimates of the benchmark model with constant parameters (data ranges from 01-01-2011 until 31-08-2021).

	Parameter estimates per company of ARMA-apARCH model											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
ARMA parameters												
μ	0.165	0.234	-0.023	0.025	0.0	0.079**	0.192***	0.211**	0.149***	0.001**	0.584	0.068***
	(1.5e-01)	(2.1e-01)	(4.4e-02)	(1.4e-01)	(3.5e-02)	(3.8e-02)	(7.1e-03)	(1.1e-01)	(4.9e-02)	(4.4e-04)	(7.6e-01)	(2.6e-02)
γ_1	-0.408	-0.67	-0.955***	-0.612***	-0.707***	-0.69***	-0.983***	0.557***	-0.674***	0.986***	-1.0***	-0.471***
	(4.4e-01)	(7.6e-01)	(1.1e-03)	(7.5e-02)	(8.5e-03)	(9.7e-03)	(1.2e-03)	(4.0e-04)	(1.8e-02)	(2.6e-04)	(1.3e-05)	(4.5e-02)
γ_2		-0.007						-0.816***	-0.081***			-0.083***
		(3.2e-01)						(2.5e-02)	(2.8e-02)			(2.8e-02)
γ_3									-0.055**			
									(2.5e-02)			
δ_1	0.436***	0.669***	0.947***	0.59***	0.686***	0.682***	0.933***	-0.572***	0.626***	-0.995***	1.0	0.414***
	(3.3e-04)	(4.1e-03)	(2.3e-04)	(3.6e-02)	(6.1e-05)	(3.7e-03)	(1.0e-03)	(2.7e-02)	(7.8e-05)	(9.7e-03)	(2.0e+00)	(9.4e-06)
δ_2	0.021						-0.058***	0.855***			-0.004	
	(1.2e-01)						(9.5e-05)	(2.7e-03)			(1.2e-02)	
δ_3	-0.034											
	(2.7e-02)											
apARCH parameters												
ω	0.2***	0.417***	0.047***	0.098***	0.037***	0.154***	0.314***	0.226***	0.277***	0.217***	0.356***	0.374***
	(4.6e-26)	(5.7e-06)	(5.6e-08)	(3.3e-06)	(3.8e-08)	(3.6e-06)	(3.0e-06)	(1.1e-07)	(1.6e-26)	(3.7e-13)	(1.5e-03)	(2.8e-05)
β	0.829***	0.757***	0.896***	0.814***	0.879***	0.728***	0.708***	0.947***	0.795***	0.871***	0.923***	0.516***
	(2.6e-06)	(3.3e-03)	(5.4e-06)	(5.0e-02)	(1.1e-05)	(3.9e-03)	(7.6e-05)	(3.5e-04)	(1.9e-06)	(1.2e-02)	(6.9e-07)	(3.1e-04)
α	0.055*	0.143***	0.075***	0.084***	0.102***	0.112***	0.162***	0.029***	0.054***	0.044***	0.047***	0.242***
	(4.0e-02)	(3.9e-02)	(3.5e-04)	(6.7e-04)	(8.2e-04)	(4.5e-03)	(1.6e-03)	(8.5e-04)	(3.7e-03)	(3.8e-03)	(5.6e-09)	(2.3e-04)
ψ	1.0**	0.188	0.36***	0.332***	0.164***	0.526***	0.201**	0.42**	1.0***	1.0***	-0.013	0.054
	(5.6e-01)	(1.5e-01)	(7.7e-02)	(1.1e-01)	(5.2e-02)	(1.0e-01)	(9.4e-02)	(2.1e-01)	(2.8e-01)	(1.1e-01)	(5.3e-01)	(1.2e-01)

Parameter estimates per company of the ARMA-apARCH benchmark model. Table indicates the parameters of the conditional mean equation and the conditional variance equation. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters (tested using equation (35)) at the 10%, 5% and 1% significance level, respectively.

A.1.4 Time-varying parameter estimates

Here, the parameter estimates of the various models that allow for time-varying parameters in the volatility equation are presented.

Table 15 – Parameter estimates of Model I with time-varying conditional volatility parameters (data ranges from 01-01-2011 until 31-08-2021).

	Parameter estimates per company of the vpARMAX-apARCH model											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
ARMAX parameters												
μ	0.165***	0.232***	0.0001	0.04*	-0.00034	0.088***	0.187***	0.00748***	0.148**	-0.00011***	0.084**	0.08***
	(4.1e-02)	(9.1e-02)	(1.3e-04)	(2.7e-02)	(3.9e-02)	(3.5e-02)	(3.6e-02)	(2.0e-03)	(6.5e-02)	(1.5e-06)	(4.0e-02)	(2.9e-02)
γ_1	0.027	0.017	0.00039	0.017	-0.021	0.00704	-0.00352	-0.013	0.039	0.00509***	0.013	-0.00124
	(2.9e-02)	(4.9e-02)	(9.2e-04)	(1.7e-02)	(3.3e-02)	(1.6e-02)	(3.5e-02)	(4.5e-02)	(3.3e-02)	(7.7e-06)	(4.0e-02)	(3.4e-02)
γ_2		-0.045*						0.017	-0.021			0.00333
		(3.2e-02)						(4.3e-02)	(3.1e-02)			(2.1e-02)
γ_3									0.08***			
									(3.3e-02)			
δ_1	-0.051*	-0.00754	-0.00168	0.012	-0.032	0.022**	0.00793	-0.00369	-0.628***	-0.00193***	-0.048	0.039**
	(3.3e-02)	(8.2e-02)	(1.9e-03)	(1.8e-02)	(2.9e-02)	(1.3e-02)	(3.7e-02)	(4.2e-03)	(5.7e-02)	(2.7e-05)	(7.4e-02)	(1.9e-02)
δ_2	-0.045						-0.026**	-0.015**			-0.0006	
	(4.7e-02)						(1.1e-02)	(8.2e-03)			(6.3e-02)	
δ_3	-0.397***											
	(3.2e-02)											
$\pi_{sent,1}$	0.418***	0.00374	0.00042	0.017	0.00327	-0.00634	0.00722*	0.019	-0.068**	-0.00191***	0.367***	-0.485***
	(1.1e-05)	(3.7e-02)	(1.0e-03)	(1.4e-02)	(2.9e-02)	(1.3e-02)	(5.6e-03)	(1.8e-02)	(3.6e-02)	(4.0e-05)	(4.3e-02)	(5.0e-02)
$\pi_{sent,2}$		-0.773**			-0.029		-0.98***	0.00988				
		(3.6e-01)			(4.2e-02)		(7.8e-02)	(1.3e-02)				
$\pi_{tweets,1}$	0.015	-0.00666	0.984***	-0.512***	-0.883***	-0.687***	0.936***	-0.00863**	-0.047***	0.996***	-0.378***	-0.071***
	(2.3e-02)	(1.5e-01)	(4.7e-04)	(2.8e-02)	(2.7e-01)	(5.7e-02)	(1.6e-01)	(5.0e-03)	(1.7e-02)	(4.6e-05)	(2.3e-05)	(2.4e-02)
$\pi_{tweets,2}$								0.964***				
								(5.3e-03)				
$\pi_{interact,1}$	-0.044**	0.77***	-0.996***	0.481***	0.876***	0.672***	-0.057***	-0.00228	0.589***	-1.003	-0.00784	0.438***
	(2.3e-02)	(6.2e-02)	(2.2e-03)	(3.4e-05)	(1.1e-03)	(1.2e-04)	(1.1e-04)	(9.6e-03)	(9.7e-05)	(4.5e+00)	(2.7e-02)	(2.8e-05)
$\pi_{interact,2}$								-0.96***				
								(6.2e-04)				
Time-varying apARCH parameters												
ω	0.213***	0.743***	0.054***	0.212***	0.037***	0.111***	0.396***	1.982***	1.488***	0.055***	0.367***	0.677***
	(2.8e-09)	(1.6e-08)	(1.3e-06)	(1.8e-10)	(7.3e-07)	(2.3e-08)	(6.6e-06)	(7.3e-06)	(3.0e-07)	(7.7e-04)	(5.2e-04)	(1.7e-07)
ω^*	-0.031	-0.45***	0.00852	-0.08***	0.00782	0.25***	-0.154***	-1.841***	-1.233***	-0.014	-0.355***	-0.1
	(5.3e-02)	(3.6e-04)	(1.4e-02)	(5.9e-04)	(1.4e-02)	(3.7e-03)	(1.8e-03)	(8.1e-03)	(2.1e-03)	(1.0e-01)	(4.7e-04)	(9.6e-02)
β	0.818***	0.684***	0.911***	0.685***	0.883***	0.794***	0.676***	0.564***	0.098***	0.927***	0.923***	0.00139***
	(2.3e-04)	(5.6e-02)	(5.5e-04)	(2.2e-04)	(7.1e-04)	(1.2e-04)	(1.1e-01)	(5.3e-04)	(5.9e-03)	(2.0e-05)	(1.9e-03)	(5.4e-04)
β^*	-0.012	0.093***	-0.033***	0.102***	-0.00723	-0.245***	0.08***	0.394***	0.727***	-0.012	0.062***	0.533***
	(1.7e-02)	(2.6e-02)	(6.2e-03)	(6.6e-03)	(3.1e-02)	(5.1e-03)	(2.3e-02)	(3.8e-04)	(1.0e-03)	(2.5e-02)	(1.1e-03)	(8.9e-04)
α	0.084***	0.056***	0.047***	0.019***	0.093***	0.025***	0.112***	0.158***	0.091***	0.058***	0.046***	0.357***
	(1.5e-08)	(7.1e-06)	(1.6e-05)	(1.7e-07)	(7.4e-07)	(2.4e-03)	(6.2e-04)	(1.6e-04)	(9.6e-05)	(3.5e-05)	(6.7e-06)	(7.2e-04)
α^*	0.044***	0.133***	0.018***	0.102***	0.00982	0.217***	0.031	-0.128***	-0.036***	-0.022***	-0.034***	-0.126***
	(4.8e-03)	(1.7e-03)	(3.0e-03)	(1.3e-03)	(3.2e-02)	(2.4e-02)	(4.6e-02)	(6.8e-05)	(1.1e-03)	(5.9e-04)	(4.5e-05)	(9.2e-03)
ψ	0.471***	-0.123	0.266***	1.0***	0.067	1.0***	-0.1	-0.45***	0.98**	0.521	-0.048	0.201
	(1.2e-01)	(9.9e-01)	(4.7e-03)	(9.9e-02)	(7.4e-02)	(1.2e-01)	(3.0e-01)	(3.8e-02)	(4.3e-01)	(1.2e+00)	(2.9e-01)	(2.0e-01)
ψ^*	-0.022	0.305	0.476***	-0.734***	0.197	-0.619***	0.505	1.0***	0.02	0.479	0.311**	-0.15
	(4.6e-02)	(1.1e+00)	(1.6e-01)	(1.6e-01)	(1.6e-01)	(2.2e-01)	(4.3e-01)	(3.1e-01)	(3.3e-02)	(3.5e+00)	(1.7e-01)	(2.2e-01)

Time-varying parameter estimates per company of the ARMAX-apARCH model. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1% significance level, respectively.

Table 16 – Parameter estimates of Model II with time-varying conditional volatility parameters (data ranges from 01-01-2011 until 31-08-2021). Table continues on next page.

	Parameter estimates per company of the vpARMA-apARCH-apX model											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
ARMA parameters												
μ	0.191***	0.233***	-0.019	0.035**	0.00183	0.086***	0.179***	0.264	0.127***	0.00347***	0.162**	0.033
	(4.3e-02)	(5.8e-02)	(6.9e-02)	(1.7e-02)	(3.6e-02)	(3.0e-02)	(3.2e-02)	(2.6e+00)	(4.5e-02)	(8.9e-04)	(7.2e-02)	(3.1e-02)
γ_1	-0.453***	-0.744***	-0.896***	-0.576***	-0.71***	-0.692***	-0.98***	-0.926	-0.548***	0.961***	-0.197***	-0.502***
	(4.2e-02)	(5.5e-02)	(1.2e-01)	(3.4e-02)	(8.1e-02)	(6.5e-02)	(6.4e-02)	(1.3e+00)	(1.6e-01)	(1.3e-02)	(4.6e-02)	(6.3e-02)
γ_2		-0.016						0.00742	-0.056			-0.039*
		(3.6e-02)						(1.3e+01)	(1.4e-01)			(2.9e-02)
γ_3									-0.042			
									(8.5e-02)			
δ_1	0.463***	0.733***	0.882***	0.554***	0.692***	0.677***	0.934***	0.932***	0.51***	-0.983***	0.186***	0.453***
	(4.0e-05)	(6.5e-05)	(5.7e-04)	(6.0e-04)	(1.4e-04)	(1.4e-03)	(2.7e-02)	(6.0e-03)	(3.4e-03)	(3.0e-02)	(2.3e-02)	(1.3e-04)
δ_2	0.011						-0.055***	0.00345			-0.00984	
	(5.7e-02)						(2.2e-04)	(1.3e+01)			(2.3e-02)	
δ_3	-0.042											
	(5.3e-02)											
Time-varying apARCH-apX parameters												
ω	0.067***	0.109***	0.026***	0.09***	0.036***	0.109***	0.215***	0.018***	1.304***	0.295***	0.236***	0.00466***
	(9.2e-03)	(6.7e-04)	(8.5e-06)	(4.1e-08)	(1.1e-04)	(3.3e-06)	(6.8e-06)	(2.0e-07)	(8.2e-10)	(4.7e-03)	(2.8e-03)	(4.6e-05)
ω^*	-0.021	0.056***	0.015*	0.00546	-0.036***	0.198***	-0.167***	-0.018***	-1.198***	-0.295***	0.644***	0.303***
	(1.9e-02)	(1.1e-02)	(1.2e-02)	(2.2e-02)	(4.5e-03)	(4.2e-03)	(1.7e-03)	(4.9e-07)	(1.7e-02)	(4.1e-03)	(2.6e-03)	(3.8e-03)
β	0.828***	0.881***	0.914***	0.837***	0.886***	0.787***	0.753***	0.985***	0.192***	0.615***	0.923***	0.966***
	(1.8e-03)	(8.5e-06)	(5.2e-05)	(8.8e-04)	(1.5e-04)	(1.7e-03)	(2.1e-02)	(3.3e-03)	(6.0e-03)	(1.9e-02)	(6.6e-07)	(2.2e-04)
β^*	0.00357	-0.125***	-0.034***	-0.041***	-0.028*	-0.225***	0.081***	-0.018	0.557***	0.244***	-0.126***	-0.231***
	(1.6e-02)	(4.8e-03)	(5.3e-03)	(6.9e-03)	(1.7e-02)	(5.5e-03)	(1.0e-02)	(2.4e-01)	(1.2e-02)	(7.2e-03)	(1.1e-03)	(1.3e-03)
α	0.06***	0.0	0.044***	0.017***	0.087***	0.027***	0.065***	0.0	0.077***	0.071***	0.041***	0.0
	(3.4e-03)	(1.2e-04)	(5.8e-04)	(7.9e-04)	(2.6e-04)	(2.2e-03)	(1.3e-03)	(6.0e-04)	(2.0e-05)	(8.1e-05)	(1.7e-04)	(1.6e-04)
α^*	0.046***	0.193***	0.048***	0.103***	0.029*	0.205***	0.015	0.016	-0.015**	-0.023***	0.068***	0.091***
	(1.6e-02)	(6.0e-03)	(4.5e-04)	(3.6e-04)	(2.0e-02)	(1.3e-02)	(3.7e-02)	(5.6e-02)	(7.3e-03)	(2.0e-03)	(5.0e-04)	(1.4e-02)
ψ	0.443***	0.291	0.232**	1.0***	0.033	1.0***	-0.341**	-0.241	1.0	0.98***	-0.122	-0.586**
	(1.3e-01)	(4.0e-01)	(1.1e-01)	(1.4e-01)	(1.2e-01)	(1.6e-01)	(2.1e-01)	(1.1e+01)	(1.4e+00)	(1.9e-01)	(1.3e-01)	(3.5e-01)
ψ^*	0.1	-0.152	0.253*	-0.736***	0.243	-0.614***	0.995**	1.0	2e-05	0.02	0.195	0.999***
	(1.6e-01)	(3.5e-01)	(1.6e-01)	(2.1e-01)	(2.1e-01)	(2.1e-01)	(5.2e-01)	(3.2e+00)	(7.3e-03)	(7.6e-02)	(1.7e-01)	(2.9e-01)
$\pi_{sent,1}$	0.014	0.00594	0.0	0.00086	0.0	0.00368	0.062***	0.00446	0.0	0.335***	0.035***	0.00536***
	(1.8e-02)	(3.4e-02)	(1.8e-05)	(3.3e-03)	(1.2e-04)	(9.2e-03)	(5.8e-04)	(2.7e-01)	(1.0e-04)	(1.7e-02)	(9.1e-03)	(1.8e-03)
$\pi^*_{sent,1}$	0.033***	-0.00584	0.0001	-0.00076	0.00455	-0.00358	-0.062***	0.072	0.105**	-0.219***	-0.035	-0.00526
	(7.6e-03)	(3.1e-02)	(1.2e-04)	(3.0e-03)	(9.6e-03)	(7.6e-03)	(2.8e-03)	(5.8e-01)	(4.8e-02)	(6.7e-02)	(3.1e-02)	(1.3e-01)
$\pi_{sent,2}$							0.0				0.0	
							(2.8e-05)				(4.8e-06)	
$\pi^*_{sent,2}$							0.054***				0.0001***	
							(5.1e-03)				(2.7e-05)	
$\pi_{tweets,1}$	0.153***	0.08***	0.251***	0.0	0.00084***	0.0	0.0	0.222***	0.0	0.043	0.148***	0.036***
	(1.7e-03)	(2.6e-02)	(3.5e-03)	(1.1e-04)	(1.6e-04)	(3.6e-04)	(7.2e-05)	(9.0e-03)	(9.6e-05)	(5.2e-02)	(5.4e-02)	(1.1e-03)
$\pi^*_{tweets,1}$	-0.153***	0.117	-0.236***	0.00842***	-0.00074**	0.127***	0.031***	-0.189	0.606***	0.059	0.444***	-0.024
	(1.6e-03)	(1.2e-01)	(1.2e-02)	(4.4e-04)	(4.4e-04)	(4.5e-02)	(2.7e-03)	(2.7e-01)	(1.5e-02)	(9.3e-02)	(1.7e-01)	(4.9e-02)
$\pi_{interact,1}$	0.487***	1.761***	0.00103***	0.0	0.00244***	0.0	0.065	1.0***	0.0	1.188***	0.011***	0.0
	(1.3e-02)	(9.5e-02)	(1.5e-04)	(9.2e-03)	(7.5e-04)	(2.9e-05)	(3.5e-01)	(1.5e-01)	(1.2e-04)	(5.4e-02)	(4.5e-03)	(1.4e-05)
$\pi^*_{interact,1}$	-0.482***	-1.0	0.023***	0.803***	2.274***	0.0001	-0.031	-1.0	0.0001	-1.0***	-0.011***	0.0001
	(4.3e-04)	(2.2e+00)	(7.4e-03)	(7.9e-02)	(6.9e-02)	(5.8e-04)	(6.4e-01)	(1.3e+00)	(2.6e-04)	(4.0e-02)	(3.9e-03)	(1.0e-03)

	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
$\psi_{sent,1}$	1.0	0.625	-0.25	0.943	-0.933	0.811	-0.97***	-0.387	-0.354	0.197*	0.927**	0.516
	(8.8e-01)	(7.4e+00)	(1.6e+00)	(3.1e+00)	(4.3e+00)	(1.1e+00)	(2.4e-02)	(8.3e-01)	(3.0e+00)	(1.3e-01)	(4.6e-01)	(6.3e-01)
$\psi^*_{sent,1}$	-0.035	-1.0	0.453	-0.111	-0.064	-1.0	0.517***	1.0	1.0	-0.197	-0.396	-1.0
	(2.0e-01)	(2.2e+00)	(2.1e+00)	(2.7e+00)	(4.2e+00)	(1.4e+00)	(2.1e-01)	(1.7e+01)	(2.3e+00)	(2.2e-01)	(2.2e+00)	(2.7e+01)
$\psi_{sent,2}$							-0.207				-0.944	
							(3.4e-01)				(9.8e-01)	
$\psi^*_{sent,2}$							0.703***				-0.047	
							(2.3e-01)				(1.1e+00)	
$\psi_{tweets,1}$	0.365*	0.163***	0.894***	-0.567	-0.22	0.961	-0.684	0.838	1.0***	0.987***	0.994***	0.864***
	(2.8e-01)	(3.4e-03)	(3.3e-01)	(7.8e-01)	(7.9e-01)	(8.9e-01)	(1.5e+00)	(3.1e+00)	(4.1e-01)	(1.8e-01)	(9.2e-02)	(2.2e-01)
$\psi^*_{tweets,1}$	0.635	0.837***	0.106	0.656	1.0	-0.504	1.0	0.16	-0.435**	0.013	0.00632	-1.0
	(8.2e-01)	(2.5e-01)	(3.3e-01)	(6.8e-01)	(1.0e+00)	(5.5e-01)	(1.8e+00)	(6.3e-01)	(2.0e-01)	(1.1e-02)	(1.6e-01)	(2.1e+00)
$\psi_{interact,1}$	0.693***	0.999*	-0.777	1.0	-1e-05	0.957	1.0	0.963	-0.986	0.545*	0.033	0.58***
	(2.4e-01)	(7.3e-01)	(9.9e-01)	(8.1e-01)	(1.4e-01)	(3.2e+01)	(1.8e+00)	(1.3e+01)	(1.2e+00)	(3.3e-01)	(5.7e-01)	(8.9e-02)
$\psi^*_{interact,1}$	-1.0	-0.144	-0.223	-1.0*	1.0***	-0.062	-0.506	-0.452	0.016	-0.021	0.534	0.293
	(3.0e+00)	(1.4e+00)	(3.0e+00)	(7.1e-01)	(3.0e-01)	(3.2e+01)	(3.1e+00)	(9.6e+00)	(3.9e+00)	(2.2e-02)	(4.8e-01)	(8.0e-01)

Time-varying parameter estimates per company of the ARMA-apARCH-apX model. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1% significance level, respectively.

Table 17 – Parameter estimates of Model III with time-varying conditional volatility parameters (data ranges from 01-01-2011 until 31-08-2021). Table continues on next page.

	Parameter estimates per company of the vpARMA-apARCH-apXGARCH model											
	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
ARMA parameters												
μ	0.176**	0.236***	-0.01	0.04	0.00323	0.087	0.187***	0.277*	0.138***	0.147	0.565	0.086***
	(8.9e-02)	(6.5e-02)	(5.9e-02)	(3.3e-02)	(3.5e-02)	(1.9e-01)	(4.4e-02)	(1.8e-01)	(2.9e-02)	(5.8e-01)	(7.1e-01)	(3.0e-02)
γ_1	-0.426***	-0.673***	-0.889***	-0.56***	-0.707***	-0.695***	-0.98***	-0.857***	-0.6***	-0.662*	-1.0***	-0.548***
	(3.3e-02)	(1.8e-01)	(1.3e-01)	(3.4e-02)	(9.6e-02)	(1.0e-01)	(1.6e-03)	(2.3e-01)	(1.1e-01)	(4.8e-01)	(1.2e-02)	(5.2e-02)
γ_2		-0.00389						0.113***	-0.077***			-0.064***
		(6.6e-02)						(2.7e-02)	(3.2e-02)			(2.1e-02)
γ_3									-0.055***			
									(2.2e-02)			
δ_1	0.444***	0.673***	0.875***	0.531***	0.689***	0.68***	0.93***	0.846***	0.563***	0.627***	1.0	0.504***
	(1.9e-05)	(8.9e-03)	(7.7e-04)	(1.8e-03)	(3.2e-04)	(5.7e-04)	(4.1e-03)	(1.2e-03)	(2.9e-03)	(9.5e-04)	(1.5e+00)	(4.2e-04)
δ_2	0.016						-0.058***	-0.114***			-0.0034	
	(2.3e-02)						(9.0e-05)	(3.6e-04)			(3.4e-03)	
δ_3	-0.04**											
	(2.2e-02)											
Time-varying apARCH-apXGARCH parameters												
ω	0.175***	0.374***	0.055***	0.124***	0.034***	0.101***	0.142***	4.978***	0.744***	0.771***	0.364***	0.119***
	(1.5e-03)	(2.5e-05)	(6.2e-05)	(2.7e-07)	(1.2e-06)	(4.3e-05)	(7.4e-05)	(2.0e-04)	(1.5e-04)	(1.6e-02)	(3.7e-04)	(6.5e-04)
ω^*	-0.174***	-0.159***	-0.015	-0.05***	-0.012	-0.101**	-0.142***	-4.978***	-0.601***	-0.628***	-0.355***	0.094***
	(2.2e-03)	(8.0e-03)	(1.4e-02)	(1.4e-02)	(2.4e-02)	(5.9e-02)	(3.5e-03)	(4.5e-01)	(1.2e-04)	(2.9e-03)	(5.7e-03)	(1.1e-03)
β	0.848***	0.801***	0.906***	0.779***	0.895***	0.778***	0.878***	0.149***	0.301***	0.553***	0.924***	0.0
	(3.0e-04)	(1.1e-02)	(4.3e-04)	(5.1e-04)	(3.6e-04)	(5.7e-04)	(3.5e-03)	(1.3e-02)	(6.0e-04)	(1.0e-03)	(5.1e-04)	(1.8e-05)
β^*	-0.064***	0.03	-0.027***	-0.047*	-0.015*	-0.228***	-0.088***	0.81***	0.524***	0.333***	0.062***	0.35***
	(9.9e-03)	(3.2e-02)	(5.7e-03)	(3.6e-02)	(1.1e-02)	(1.9e-02)	(1.1e-02)	(3.6e-04)	(1.1e-02)	(1.5e-02)	(5.2e-03)	(5.9e-03)
α	0.045***	0.063***	0.047***	0.022***	0.079***	0.029***	0.02***	0.188***	0.087***	0.083***	0.045***	0.291***
	(2.9e-03)	(1.5e-03)	(2.0e-03)	(1.4e-03)	(5.5e-06)	(4.2e-03)	(8.6e-04)	(1.0e-04)	(4.3e-05)	(1.7e-03)	(5.0e-03)	(1.5e-03)
α^*	0.059***	0.066***	0.039***	0.128***	0.02***	0.213***	0.051***	-0.155***	-0.035***	-0.038	-0.033***	-0.022
	(9.2e-04)	(7.6e-03)	(7.1e-03)	(2.6e-03)	(8.0e-03)	(2.2e-02)	(1.3e-03)	(2.2e-06)	(1.2e-03)	(3.2e-02)	(2.7e-04)	(1.0e-01)
ψ	0.768*	-0.224	0.214**	1.0***	0.015	1.0***	-0.236	-0.465	0.991	0.999	-0.079	0.199
	(5.7e-01)	(2.1e+00)	(1.1e-01)	(1.3e-02)	(7.4e-02)	(3.9e-01)	(2.1e-01)	(4.2e-01)	(9.8e-01)	(5.2e+00)	(1.0e-01)	(2.5e-01)
ψ^*	-0.137	0.5	0.289**	-0.808***	0.268***	-0.614	1.0***	1.0*	0.00875	-0.00335	0.313	-0.133
	(6.5e-01)	(2.8e+00)	(1.7e-01)	(3.0e-01)	(1.0e-01)	(5.8e-01)	(2.6e-01)	(7.3e-01)	(7.4e-03)	(2.9e+00)	(3.5e+00)	(3.4e-01)

	AAPL	AMZN	CVX	KO	XOM	MCD	MSFT	NFLX	NKE	CRM	TSLA	WMT
$\pi_{sent,1}$	0.0	0.0	0.0	0.0	0.222*	0.0	0.0	0.0	0.0	0.683***	0.0	0.0
	(6.9e-05)	(1.0e-04)	(4.5e-05)	(1.4e-05)	(1.5e-01)	(2.2e-04)	(8.5e-04)	(4.1e-05)	(9.5e-06)	(4.5e-02)	(8.1e-04)	(4.3e-05)
$\pi^*_{sent,1}$	0.00202***	0.0001	0.0001**	0.012	-0.053	0.00725	0.00644***	0.012***	0.578**	-0.523*	0.0001	0.0001
	(1.1e-05)	(2.3e-03)	(5.6e-05)	(2.1e-02)	(2.7e-01)	(4.9e-02)	(9.8e-04)	(1.4e-03)	(3.1e-01)	(3.8e-01)	(2.6e-03)	(1.1e-03)
$\pi_{sent,2}$		0.0										
		(1.4e-04)										
$\pi^*_{sent,2}$		0.0001										
		(1.5e-03)										
$\pi_{tweets,1}$	0.00187***	0.017***	0.00029***	0.0	9e-05	0.0	0.0	0.0	0.00013	0.00155	1e-05	0.0
	(3.1e-04)	(1.1e-03)	(9.2e-07)	(1.1e-06)	(2.6e-04)	(7.0e-05)	(4.1e-04)	(3.4e-05)	(1.7e-02)	(8.5e-03)	(7.1e+00)	(1.1e-08)
$\pi^*_{tweets,1}$	-0.00177	-0.016	-0.00018***	0.0001***	1e-05	0.0001	0.036***	0.00032**	0.05***	-0.00145	0.00069	0.0001***
	(1.9e-03)	(1.4e-01)	(6.3e-05)	(3.4e-05)	(2.7e-03)	(2.4e-04)	(5.2e-03)	(2.0e-04)	(2.1e-02)	(4.2e-03)	(7.5e+00)	(4.0e-05)
$\pi_{interact,1}$	0.0	0.0	0.00043	0.0	0.0	0.0	0.0	0.0	0.00779	0.011	0.0	0.00153*
	(3.0e-05)	(5.5e-05)	(4.0e-04)	(2.3e-05)	(2.0e-05)	(2.5e-06)	(3.8e-06)	(5.4e-08)	(8.2e-02)	(1.0e-02)	(1.2e-03)	(1.1e-03)
$\pi^*_{interact,1}$	0.0003	0.00017	-0.00033	0.0001***	0.098***	0.00036	0.0001	0.0001	0.102	-0.00716	0.0001	0.0
	(2.7e-04)	(3.2e-04)	(7.5e-03)	(1.7e-05)	(4.2e-02)	(7.9e-04)	(1.5e-03)	(2.4e-04)	(2.3e-01)	(2.5e-02)	(9.4e-02)	(8.1e-01)
$\lambda_{sent,1}$	0.0	0.0	0.0	0.0	0.0	0.018*	0.0	0.0	0.0	0.0	0.0	0.513***
	(7.7e-05)	(3.3e-04)	(2.0e-05)	(8.0e-06)	(6.2e-05)	(1.2e-02)	(1.4e-03)	(4.5e-02)	(1.1e-05)	(9.2e-06)	(4.1e-03)	(7.8e-03)
$\lambda^*_{sent,1}$	0.0001	0.0001	0.0001	0.0001	0.0001	0.139	0.177***	0.0001	0.0001	0.0001	0.0001	-0.513***
	(2.1e+00)	(2.6e-01)	(1.7e-02)	(6.2e-02)	(4.7e-02)	(3.9e-01)	(2.6e-03)	(6.4e-01)	(4.6e-03)	(3.9e-01)	(4.8e-01)	(9.7e-04)
$\lambda_{sent,2}$	0.0											
	(3.9e-05)											
$\lambda^*_{sent,2}$	0.0001											
	(1.6e+00)											
$\lambda_{sent,3}$	0.0											
	(1.7e-04)											
$\lambda^*_{sent,3}$	0.0001											
	(1.8e+00)											
$\lambda_{tweets,1}$	0.012	0.0	0.0	0.0002	0.0	0.0	0.019***	0.0	0.16***	0.0	0.0	0.089***
	(3.3e-02)	(1.9e-04)	(2.6e-05)	(4.7e-03)	(3.2e-05)	(2.0e-04)	(7.7e-04)	(6.0e-05)	(5.8e-03)	(5.7e-01)	(6.8e-03)	(1.1e-02)
$\lambda^*_{tweets,1}$	0.128***	0.0001	0.0001	0.078***	0.0001	0.231	-0.019	0.0001	-0.16***	0.0001	0.0001	0.708***
	(5.3e-03)	(2.6e-01)	(1.2e-02)	(2.2e-02)	(3.8e-02)	(1.9e-01)	(4.2e-02)	(4.6e-01)	(2.9e-02)	(2.0e+00)	(5.2e-01)	(1.3e-02)
$\lambda_{interact,1}$	0.012	0.0	0.00394	0.0	0.0	0.0	0.0	0.0	0.291***	0.0	0.00991	0.0
	(2.5e-02)	(2.8e-04)	(3.5e-03)	(2.1e-05)	(1.1e-04)	(3.8e-05)	(1.1e-03)	(6.5e-05)	(7.8e-04)	(3.7e-06)	(6.7e-01)	(4.6e-05)
$\lambda^*_{interact,1}$	0.127***	0.00018	0.081***	0.239***	0.022	0.0001	0.151***	0.0001	-0.233***	0.0001	-0.00981	0.202***
	(1.9e-02)	(1.4e-01)	(1.4e-02)	(3.7e-03)	(1.9e-02)	(2.4e-02)	(9.1e-03)	(2.7e-01)	(1.8e-03)	(1.2e+00)	(4.0e-01)	(6.8e-02)
$\psi_{sent,1}$	-0.067	0.089	0.821	0.918***	1.0***	-0.00709	0.375	-0.408	0.144***	0.554***	-0.093	0.642
	(1.4e+00)	(1.6e+00)	(9.8e-01)	(4.8e-03)	(2.0e-08)	(9.4e-01)	(3.6e-01)	(4.3e+00)	(5.1e-02)	(5.0e-02)	(1.3e+02)	(9.2e-01)
$\psi^*_{sent,1}$	-0.933	0.911	-0.166	0.082***	-0.00013	0.661	-0.357	-0.16	0.73***	0.446	0.853	-0.154
	(1.0e+00)	(3.4e+00)	(9.3e-01)	(1.0e-02)	(1.3e-04)	(2.4e+00)	(5.4e-01)	(3.9e+00)	(1.6e-01)	(4.8e-01)	(1.3e+02)	(5.1e+00)
$\psi_{sent,2}$		0.117										
		(4.1e+00)										
$\psi^*_{sent,2}$		0.628										
		(6.3e+00)										
$\psi_{tweets,1}$	0.517***	0.763***	0.877***	0.774***	0.784	0.999*	0.094	0.602*	0.985***	1.0	1.0***	0.305
	(8.7e-02)	(1.2e-01)	(1.3e-04)	(1.0e-01)	(6.2e-01)	(6.3e-01)	(1.2e-01)	(3.7e-01)	(8.0e-02)	(1.5e+00)	(2.9e-01)	(2.6e-01)
$\psi^*_{tweets,1}$	0.483***	0.237	0.123***	-0.073	0.216	-0.338	0.858**	0.398	-0.179*	-0.754**	-0.00389	0.509**
	(1.5e-01)	(5.9e-01)	(2.5e-03)	(1.4e-01)	(6.8e-01)	(3.3e+00)	(3.9e-01)	(4.5e-01)	(1.3e-01)	(3.9e-01)	(1.8e+00)	(2.6e-01)
$\psi_{interact,1}$	0.494	0.863***	0.955***	1.0***	0.628	0.567***	0.771***	0.641***	1.0***	0.268	0.874	0.987***
	(4.9e-01)	(1.1e-01)	(3.0e-02)	(3.5e-03)	(1.5e+00)	(1.8e-01)	(2.2e-01)	(1.6e-01)	(6.2e-09)	(4.9e-01)	(8.8e+00)	(8.4e-03)
$\psi^*_{interact,1}$	0.5	-0.031	-0.158	-0.117*	0.354	0.43**	0.229	0.359**	-3e-05	0.557	0.126	0.013
	(5.2e-01)	(9.3e-01)	(5.1e-01)	(8.5e-02)	(9.8e-01)	(2.3e-01)	(8.9e-01)	(2.0e-01)	(1.2e-04)	(3.3e+00)	(9.1e+00)	(5.2e-02)

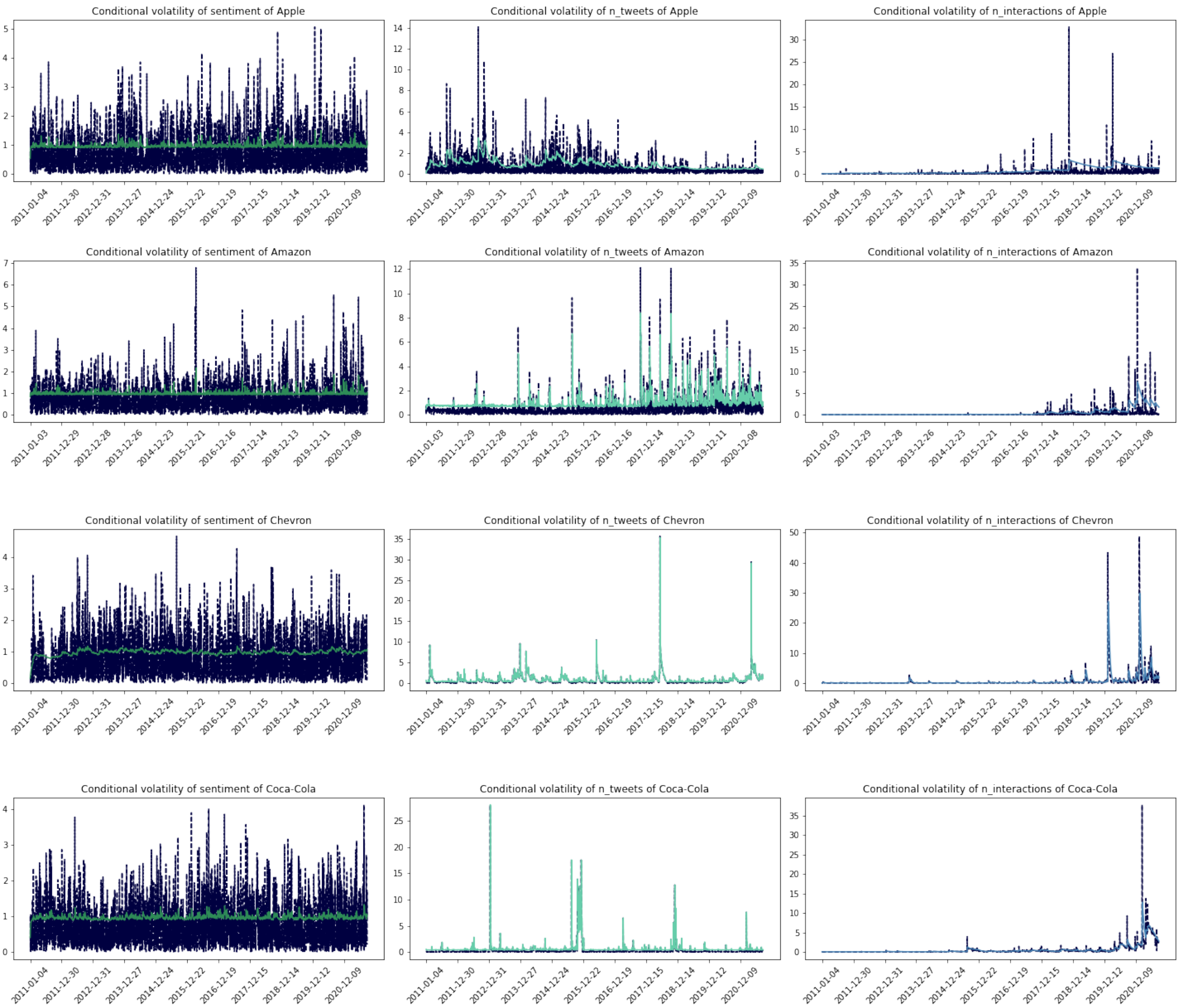
Time-varying parameter estimates per company of the ARMA-apARCH-apXGARCH model. Parentheses display the std. error of the parameter estimates. Asterisks (*, **, ***) denote significance of the parameters at the 10%, 5% and 1% significance level, respectively.

A.2 Figures

A.2.1 Plots of exogenous conditional volatility plots

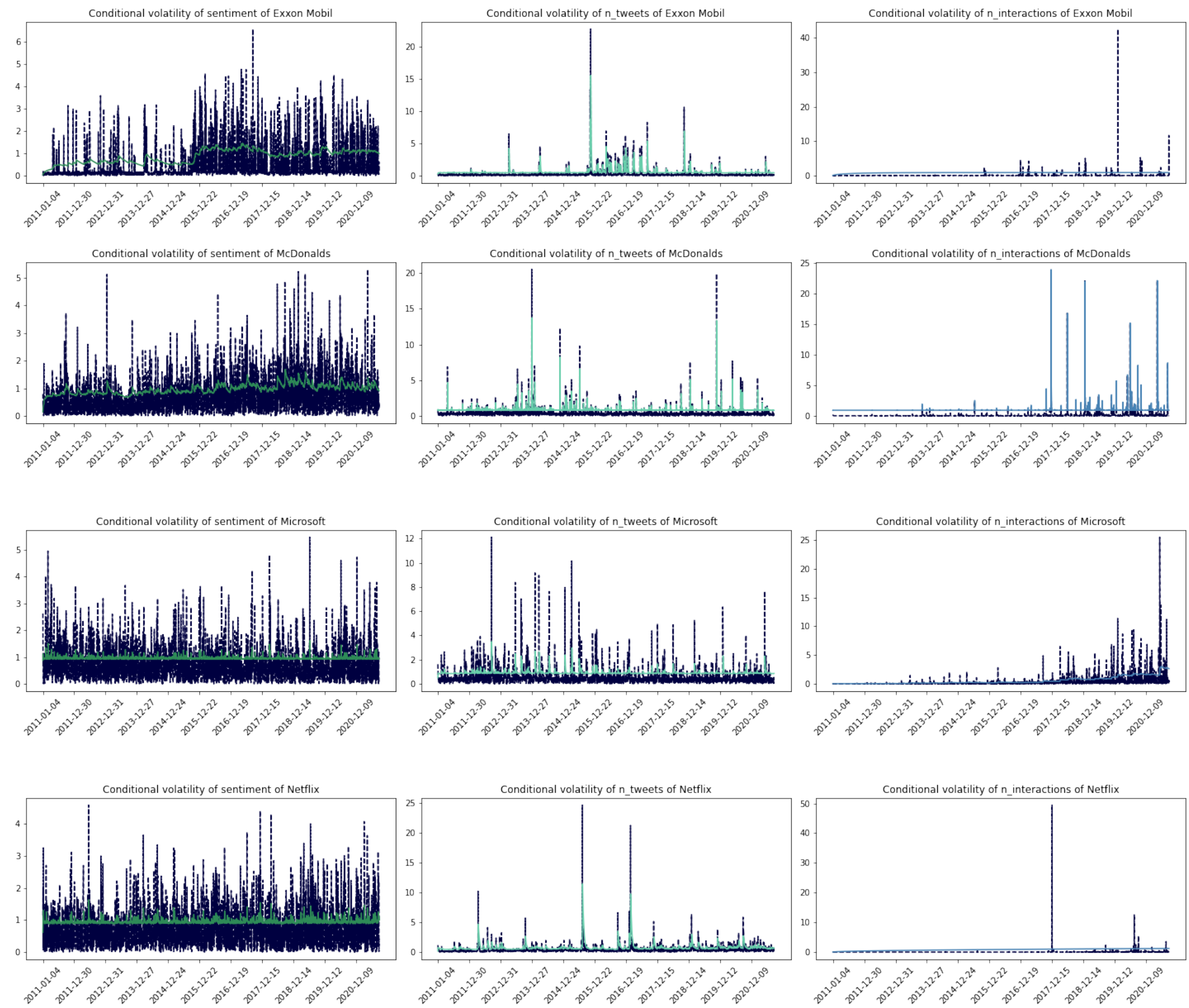
Here, the plots of the conditional volatility of the exogenous variables are given for each company included in this research.

Figure 3 – Conditional volatility of exogenous variables of U.S. companies



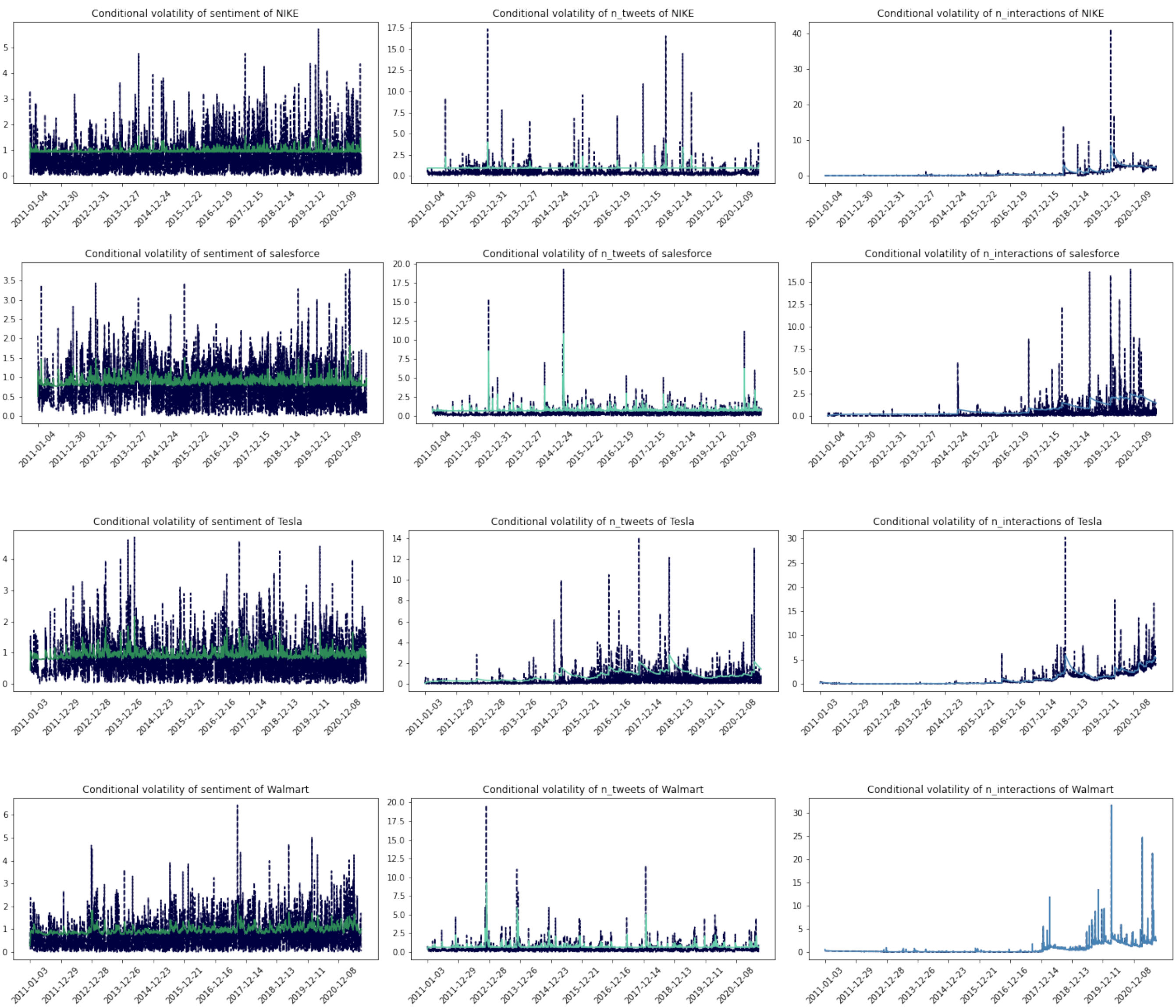
Conditional volatility for the exogenous variables of the selected U.S. companies between 01-01-2011 and 31-08-2021. Dark blue dotted line represents the square root of the squared residuals of the ARMA(1,1) process describing the mean of the exogenous variable process, colored line shows the conditional volatility $\sigma_t^{(x)}$. From left to right, plots show time series plots of the conditional volatility of Twitter sentiment, number of tweets and number of interactions of Apple, Amazon, Chevron, and Coca-Cola (up to down).

Figure 4 – Conditional volatility of exogenous variables of U.S. companies



Conditional volatility for the exogenous variables of the selected U.S. companies between 01-01-2011 and 31-08-2021. Dark blue dotted line represents the square root of the squared residuals of the ARMA(1,1) process describing the mean of the exogenous variable process, colored line shows the conditional volatility $\sigma_t^{(x)}$. From left to right, plots show time series plots of the conditional volatility of Twitter sentiment, number of tweets and number of interactions of Exxon Mobil, McDonalds, Microsoft, and Netflix (up to down).

Figure 5 – Conditional volatility of exogenous variables of U.S. companies

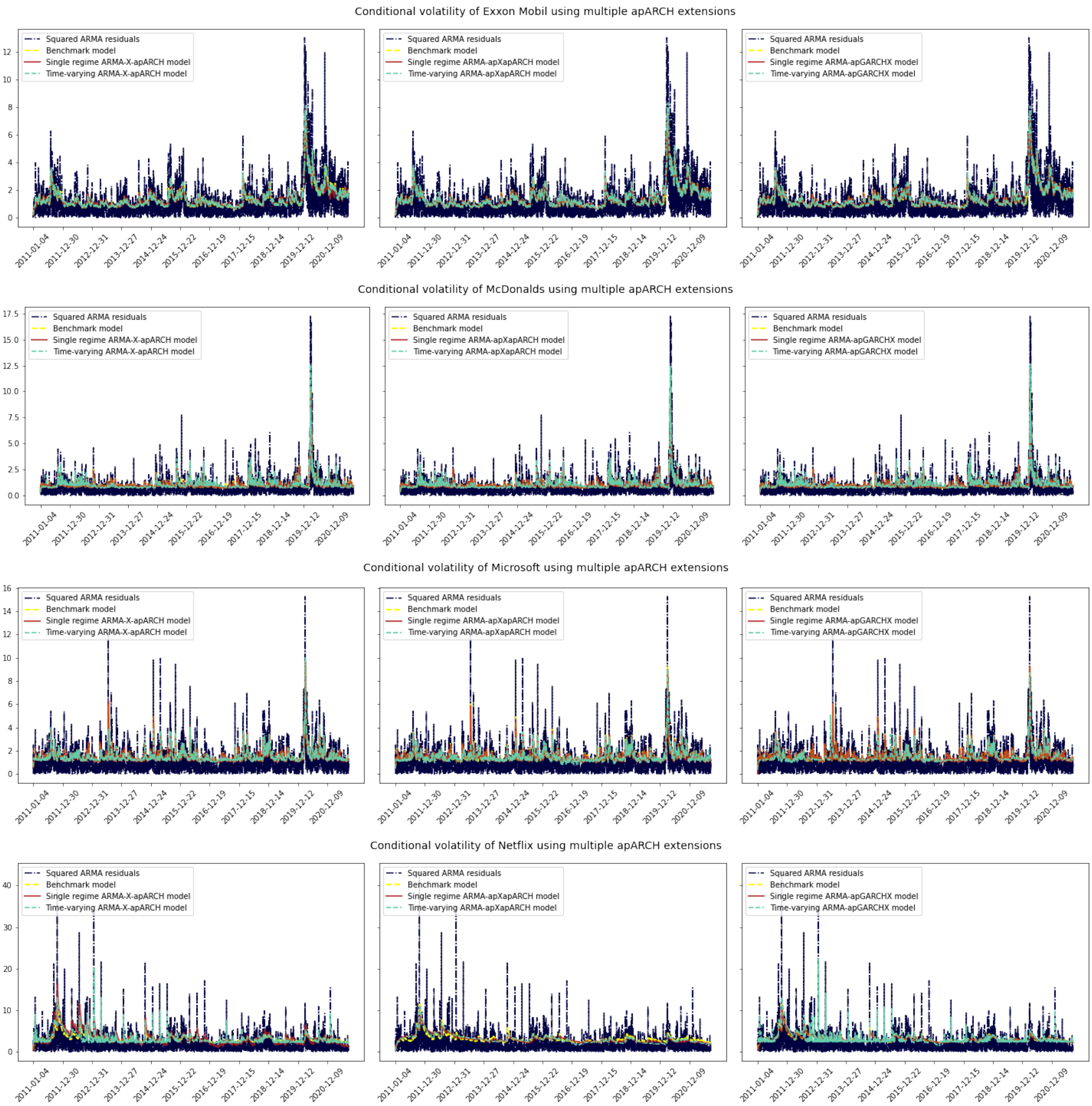


Conditional volatility for the exogenous variables of the selected U.S. companies between 01-01-2011 and 31-08-2021. Dark blue dotted line represents the square root of the squared residuals of the ARMA(1,1) process describing the mean of the exogenous variable process, colored line shows the conditional volatility $\sigma_t^{(x)}$. From left to right, plots show time series plots of the conditional volatility of Twitter sentiment, number of tweets and number of interactions of NIKE, salesforce, Tesla, and Walmart (up to down).

A.2.2 Plots of conditional volatility

This section shows the plots of the conditional volatility of the remaining companies included in this research.

Figure 6 – Conditional volatility between 01-01-2011 and 31-08-2021 of U.S. companies stock price returns.



Time series plot of the conditional volatility for Exxon Mobil, McDonalds, Microsoft, and Netflix (top to bottom plots) between 01-01-2011 and 31-08-2021. Squared ARMA residuals serves as a proxy for the conditional variance process (square root of proxy is denoted in dark blue). From left to right, plots show time series plots of the conditional volatility using the ARMAX-apARCH model, the ARMA-apARCH-apX model, and the ARMA-apARCH-apXGARCH model, for both the constant and time-varying parameter specification. In yellow, the conditional volatility of the benchmark model is plotted.

Figure 7 – Conditional volatility between 01-01-2011 and 31-08-2021 of U.S. companies stock price returns.

Time series plot of the conditional volatility for NIKE, salesforce, Tesla, and Walmart (top to bottom plots) between 01-01-2011 and 31-08-2021. Squared ARMA residuals serves as a proxy for the conditional variance process (square root of proxy is denoted in dark blue). From left to right, plots show time series plots of the conditional volatility using the ARMAX-apARCH model, the ARMA-apARCH-apX model, and the ARMA-apARCH-apXGARCH model, for both the constant and time-varying parameter specification. In yellow, the conditional volatility of the benchmark model is plotted.

References

Amado, C., & Teräsvirta, T. (2008). Modelling conditional and unconditional heteroscedasticity with smoothly time-varying structure. *SSE/EFI, Working Paper Series in Economics and Finance.*

Andreou, E., & Ghysels, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, *17*(5), 218–244.

Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, *36*, 334–357.

Bain, L. J., & Engelhardt, M. (1987). *Introduction to Probability and Mathematical Statistics - Second edition.* Duxbury Classic Series.

Barberis, N., Schleifer, A., & Vishny, R. (1998). A Model of investor sentiment. *Journal of Financial Economics*, *49*, 307–343.

Baurichter, R. (2021). Onderzoekers: aanstichter van beursuitschieters is particuliere belegger. *Het Financieele Dagblad.* https://fd.nl/beurs/1410282/aanstichter-van-beursuitschieters-is-particuliere-belegger-concluderen-economen-pik1caOavdrI

Bauwens, L., Preminger, A., & Rombouts, J. V. (2010). Theory and inference for a Markov switching GARCH model. *Econometrics Journal*, *13*, 218–244.

Black, F. (1976). Studies in Stock Price Volatility Changes. *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 177–181.

Bollersev, T., & Wooldridge, J. M. (1992). Quasi Maximum Likelihood Estimation and Inference in Dynamic Models with Time Varying Covariates. *Econometric Reviews*, *11*(2), 143–172.

Bollerslev, T. (1986). Generalzied Autoregressive Conditional Heteroskedascity. *Journal of Econometrics*, *31*, 307–327.

Bollerslev, T., & Ghysels, E. (1996). Periodic Autoregressive Conditional Heteroscedasticity. *Journal of Business & Economic Statistics*, *14*(2), 139–151.

Chen, J. (2020). TED spread. https://www.investopedia.com/terms/t/tedspread

Chen, M. P., Chen, P. F., & Lee, C. C. (2013). Asymmetric effects of investor sentiment on industry stock returns: Panel data evidence. *Emerging Markets Review*, *14*, 35–54.

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *The Journal of Finance*, *66*(5), 1461–1499.

de Winter, J., & van Dijk, D. (2021). Sentimentsindicator op basis van financieel-economisch nieuws. *ESB.*

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*, *74*(366), 427–431.

Ding, Z., Granger, C. W., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, *1*, 83–106.

Dredze, M., Kambadur, P., Kazantsev, G., Mann, G., & Osborne, M. (2016). How Twitter is Changing the Nature of Financial News Discovery. *DSMM'16: Proceedings of the Second International Workshop on Data-Science for Macro-Modeling*, *2*, 1–5.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, *50*, 987–1007.

Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance*, *1*, 237–245.

Fan, R., Talavera, O., & Tran, V. (2020). Social media, political uncertainty, and stock markets. *Review of Quantitative Finance and Accounting*, *55*, 1137–1153.

Feng, L., & Shi, Y. (2017). A simulation study on the distributions of disturbances in the GARCH model. *Cogent Economics & Finance*, *5*.

Fitzgerald, M. (2021). The GameStop mania might be over, but retail investors look here to stay. *CNBC*. https://www.cnbc.com/2021/02/13/why-retail-investors-are-here-to-stay.html

Franq, C., & Thieu, L. Q. (2019). QML Inference for Volatility Models with Covariates. *Econometric Theory*, *35*, 37–72.

Franq, C., & Zakoian, J.-M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, *10*(4), 605–637.

Franq, C., & Zakoian, J.-M. (2009). Testing the Nullity of GARCH Coefficients: Correction of the Standard Tests and Relative Efficiency Comparisons. *Journal of the American Statistical Association*, *104*(485), 313–324.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. John Wiley; Sons.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance*, *48*(5), 1779–1801.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Stanford University, technical paper.*

Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. *PSOSM '12: Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, *2*, 2–8. https://doi.org/10.1145/2185354.2185356

Han, H., & Kristensen, D. (2014). Asymptotic Theory for the QMLE in GARCH-X Models With Stationary and Nonstationary Covariates. *Journal of Business and Economic Statistics*, *3*, 167–179.

Hansen, P. R., & Lunde, A. (2005). A Forecast Comparison of Volatility Models: Does anything beat GARCH(1,1)? *Journal of Applied Econometrics*, *20*, 873–889.

Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, *131*, 97–121.

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*, 20–38.

Hsu, Y.-J., Lu, Y.-C., & Yang, J. J. (2021). News sentiment and stock market volatility. *Review of Quantitative Finance and Accounting*, *57*, 1093–1122.

Hutto, C., & Gilbert, E. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Conference: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.*

Kim, D., & Kon, S. J. (1999). Structural change and time dependence in models of stock returns. *Journal of Empirical Finance*, *6*, 283–308.

Kuepper, J. (2021). CBOE Volatility Index (VIX). https://www.investopedia.com/terms/v/vix.asp

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a Social Network or a News Media. *WWW '10: Proceedings of the 19th international conference on World wide web*, *19*, 591–600. https://doi.org/10.1145/1772690.1772751

Lamoureux, C. G., & Lastrapes, W. D. (1990). Persistence in Variance, Structural Change, and the GARCH model. *Journal of Business & Economic Statistics*, *8*(2), 225–234.

Ling, S., & McAleer, M. (2003). Asymptotic Theory for a Vector ARMA-GARCH Model. *Econometric Theory*, *19*(2), 280–310.

Ling, S., & McAleer, M. (2002). Stationarity and existence of moments of a family of GARCH processes. *Journal of Econometrics*, *106*, 109–117ch.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, *7*, 77–91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

Mittnik, S., Robinzonov, N., & Spindler, M. (2015). Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of Banking & Finance*, *58*, 1–14.

Nelson, D. B. (1991). Conditional Heteroskedascity in Asset Returns: A New Approach. *Econometrica*, *59*(2), 347–370.

Nelson, D. B. (1990). Stationarity and Persistence in the GARCH(1,1) model. *Econometric Theory*, *6*, 318–334.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*, 246–256.

Perdana, R. S., & Pinandito, A. (2018). Combining Likes-Retweet Analysis and Naive Bayes Classifier within Twitter for sentiment analysis. *Journal of Telecommunication, Electronic and Computer Engineering*, *10*, 40–46.

Rakowski, D., Shirley, S. E., & Stark, J. R. (2021). Twitter activity, investor attention, and the diffusion of information. *Financial Management*, *50*(1), 3–45.

Ramco, G., Aleksovski, D., Calderelli, G., Grcar, M., & Mozetic, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. *PLoS One*, *19*(9).

Ribeiro, F. N., Araujo, M., Goncalves, P., Goncalves, M. A., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ: Data Science*, *5*(23), 3–29.

Rupande, L., Muguto, H. T., & Muzindutsi, P.-F. (2019). Investor sentiment and stock return volatility: Evidence from the johannesburg stock exchange. *Cogent Economics & Finance*, *7*(1).

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

Schwert, G. W. (1989). Why Does Stock Market Volatility Change Over Time? *The Journal of Finance*, *44*(5), 1115–1153.

Smales, L. A. (2015). Asymmetric volatility response to news sentiment in gold futures. *Journal of International Financial Markets, Institutions & Money*, *34*, 161–172.

Smales, L. A. (2014). News sentiment and the investor fear gauge. *Finance Research Letters*, *11*, 122–130.

Thomassen, F. (2018). Estimation of a time-varying parameter GARCH model based on Google Trends. *Erasmus School of Economics Master thesis*.

Verma, M., Divya, & Sofat, S. (2014). Techniques to Detect Spammers in Twitter- A Survey. *International Journal of Computer Applications*, *85*(10), 27–32.

Weiss, A. A. (1984). ARMA models with ARCH errors. *Journal of Time Series Analysis*, *5*(2), 129–143.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, *50*(1), 1–25.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62.

Xu, S. (2018). Bayesian Naive Bayes classifiers to text classification. *Journal of Information Sciences*, *44*(1), 48–59.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, *1*, 69–90.