Using TGARCH-M to Model the Impact of Good News and Bad News on Covid-19 Related Stocks' Volatilities

Junqi Chen, Hui Li, Yan Lv

Nanjing University of Science and Technology, Nanjing, China
Email: junqiqi77_chen@foxmail.com

Abstract

In this paper, we investigate the dynamic relationship between Twitter sentiment related to vaccines and Covid-19 and the volatility of pharmaceutical stock returns. The first step is to construct a time-series Twitter sentiment index by considering the positive, negative, and neutral sentiment of tweets. A TGARCH-M model was then constructed to correlate the stock returns of five pharmaceutical companies with the Twitter sentiment. The results show that Twitter sentiment responds to stock price volatility in the market, especially in three companies, BioTech, Novovax, and Moderna. The relationship between the volatility of the stock returns of the three companies and Twitter sentiment was significant. Stock returns are negatively correlated with their volatility, with an increase in expected risk in the market leading to a corresponding decrease in returns. Positive sentiment is more likely to produce large swings in returns than negative sentiment.

Keywords

TGARCH, GARCH-M, TGARCH-M, Volatility, Sentiment Analysis

1. Introduction

This paper focuses on the impact of internet sentiment on stock volatility. In particular, Twitter sentiment related to Covid-19 and vaccines, as we need news about it in order to determine the public sentiment towards viruses and vaccines, and thus to study its relationship with stock volatility of mainstream pharmaceutical companies.

To explore this impact, we first suggest a set of reasonably valid models for portraying user sentiment indices. The literature on user sentiment is now high-

ly subjective and insufficiently normative in terms of the proxies scholars choose for investor sentiment. The variables chosen by different scholars vary due to factors such as data availability, and most do not validate the reasonableness of the variables, resulting in a wide variety of user sentiment indicators, which ultimately leads to divergent conclusions. The use of websites, social media, online message boards, and other Internet platforms to mine user sentiment has been a popular research method in recent years. This paper, therefore, proposes a new user sentiment indicator based on the fusion of user sentiment and user attention using a text mining approach. For this purpose, over two million tweets were collected on the Twitter social platform and analysed by natural language processing to obtain a sentiment score for the tweets.

The log-returns of the stocks of the five major companies were then subjected to correlation analysis, such as correlation tests, tests for smoothness, and tests for ARCH effects. It was finally concluded that the log-returns of the five pharmaceutical companies' stocks have spikes and thick tails and volatility aggregation. Leverage and asymmetry in returns were also found. That is, negative sentiment tends to induce a greater response than positive sentiment, and the impact of a decline is greater than that of an increase.

The aggregation of volatility in the log-return series is obvious. In addition, the response to user sentiment is likely to differ between stocks of different companies. This paper, therefore, uses the TGARCH-M model and adds user sentiment changes to the mean and variance equations in order to investigate the impact of user sentiment changes on the log returns and volatility of a company's stock across different company stocks.

The data used in this paper are stock data from five pharmaceutical companies (BNTX, MRNA, PFE, AZN, NVAX) and tweets related to Covid-19, vaccines. Section 2 is basically a background introduction to the research history of ARCH and GARCH-like models and the use of natural language processing methods to calculate sentiment scores for tweets. Section 3 is devoted to analysing the data we need and the tools needed to construct a user sentiment index. In Section 7, we introduce the TGARCH-M model by introducing the GARCH-like model. In Section 5, we focus on how to construct a user sentiment index. In Section 6, we perform a correlation analysis on the stock return series to set the stage for the subsequent modelling. In Section 7, we show the results of the empirical analysis. Section 8 presents future research. Section 9 presents the conclusions.

Covid-19

The 2019 outbreak of the "Covid-19", a major worldwide public health emergency, is unlike any other. It is unlike any other outbreak in terms of both its depth and its breadth of impact. Since its outbreak in 2019, the pandemic has spread rapidly around the world, with the number of cases increasing and spreading continuously, causing serious economic and social impacts worldwide. The

economy, in particular, has been very badly affected by the pandemic, leading directly to recession in some countries. The UK's GDP fell sharply in the first half of 2020, by 21.8%, the largest decline of any G7 country.

As a result of the outbreak, countries have started their own means of combating the epidemic. There is no more effective way than a blockade. City-wide, nationwide blockades began to take place in various countries. This was followed by a dramatic decline in tourism and aviation. Countries were hit hard by their own increased entry policies and the global tourism and aviation industries were hit hard.

Among these effects, the impact of the epidemic on financial markets was visible and dramatic, lasting longer than previous outbreaks, and the reaction of financial markets to the epidemic shock was dramatic and perverse. In the US financial markets, for example, there were as many as four stock meltdowns in 10 days as a result of the epidemic, compared to only five meltdowns in the history of the US stock market as at 18 March 2020. In addition to the dramatic impact on the US stock market, other global stock markets were also affected by the epidemic, with global stock markets falling by almost 30% in just 40 trading days around March 2020. On 11 March 2020 alone, stock markets in 11 countries around the world suffered meltdowns. The panic triggered by the impact of the epidemic spread around the world in a short period of time, causing massive panic in the markets and by early April 2020, 12 trillion dollars in global equity market capitalisation had evaporated as a result of the impact of the epidemic. Compare this to the impact of the SARS outbreak in 2003, which was also a "coronavirus" but had a smaller impact on global equity markets and lasted for a very short period of time. It is not comparable to the 2019 outbreak of the "Covid-19".

Large numbers of people around the world are dying from the virus, so the development of a vaccine is urgent and it is the only way to combat it. This is why countries began developing vaccines immediately after the WHO declared the coronavirus a deadly virus. Among the pharmaceutical companies developing vaccines, five companies—BioNTech, Moderna, Pfizer, AstraZeneca and Novavax, are at the forefront of vaccine development, so we use the log returns of these five companies as the subject of our study.

2. Literature Review

For a more in-depth research, we will consider the literature on stock volatility in relation to social media and news sentiment.

2.1. Efficient Market Hypothesis (EMH)

2.1.1. The Origins of the EMH

The study of asset pricing and the factors affecting price volatility can be traced back to the "random walk" hypothesis, where Louis Bachelier (1900) found that the prices of most consumer goods fluctuate randomly. That is, in a time series, the current period's product price is equal to the expectation of the next period's

product price, and the next period's product price is equal to the current period's product price plus white noise. This discovery by Louis Bachelier marked the birth of the theory of "random wandering". Roberts (1959) applied the random walk theory to the analysis of stock prices and concluded that stock prices also follow random walks. And Roberts argues that market efficiency means that asset price movements follow a random wander. Fama (1965) proposed the Efficient Market Hypothesis (EMH), which assumes that investors are able to make unbiased estimates of the information obtained and that asset prices respond to all information. According to the information about asset pricing, there are three types of efficient markets: strong, semi-strong and weak efficient markets. First, a weak efficient market is one in which asset prices reflect only the historical information relevant to the movement of asset prices. For example, historical prices, price volatility, short-term interest rates, trading volumes, etc. As a result, investors are unable to make decisions and make excess profits based on historical information. A semi-strong market is one that reflects total publicly useful information on asset prices in a timely manner, including macroeconomic conditions, company financial conditions, and product and technology conditions. Information obtained by investors from other sources has no impact on stock prices and publicly available information has no profit-making value. A strong market is one in which asset prices reflect total public and private information. All information is reflected in the price, indicating that no information can make an investment more profitable.

2.1.2. Shortcomings of the EMH

In the 1960s, the theory of the efficient market hypothesis had great success in both theoretical research and practical application, but from the 1980s onwards, scholars gradually discovered some financial phenomena that were difficult to explain using the theory of the efficient market hypothesis.

The first is the predictability of stock returns. In their study of stock portfolio returns, De Bondt and Thaler (1985) found that there is a reversal of stock market returns in the long run, i.e. a portfolio of stocks with high current returns will have worse returns in 3 - 5 years than a portfolio of stocks with low current returns. Jegadeesh and Titman (1993) found that there is "inertia" in stock prices by examining the short term movements of stocks over a period of 6 to 12 months, i.e. the stock prices tend to maintain their current upward, downward or oscillatory movements in the short term.

The second is the abnormal volatility of stock prices. On 19 October 1987, the Dow Jones Industrial Average in New York plunged 22.6% without any negative news and triggered a global stock market decline. Culter et al. (1988) analysed the 50 most volatile companies in the post-World War II period and found that the volatility was not related to whether or not information was released and what information was released. Wurgler and Zhuravskaya (2002) studied a sample of stocks selected as constituents of the S & P 500 from 1976 to 1996 and found that stocks selected as constituents of the S & P 500 increased their returns

by 3.5% compared to the period before they were selected, but that being selected as a constituent did not change the company's operations.

According to the efficient market hypothesis, stock prices only respond to new information, up or down, and the release of new information is unpredictable, so investments cannot take advantage of past price movements to achieve excess returns. However, the above research suggests that stock prices are predictable to some extent and that share price movements are not solely due to changes in a company's operations and the release of new information. This suggests that the efficient market hypothesis theory has limitations in explaining the movement of asset prices represented by equities. In order to better explain the problems associated with asset price movements, scholars have combined behavioural psychology with traditional finance, thus giving rise to behavioural finance.

2.2. Volatility

Volatility is a measure of the change in the price of a financial asset over time and it is often used to quantify the risk of a financial asset. Volatility is usually considered negative because it represents uncertainty. If a financial asset has high volatility over a period of time, this means that its price will change significantly, while low volatility means that the price will not fluctuate dramatically. In today's markets, volatility plays an crucial role in financial markets. It has also been the subject of much academic and industry attention over the past few decades. Volatility has become a key factor in many investment decisions and portfolio creation. Investors or portfolio managers also have a level of risk that they can tolerate. Having a good volatility forecast for the price of an asset holding is essential for assessing investment risk. In addition, since the Basel Accord, the management of financial risk has become more and more important. This has forced volatility analysis and forecasting to become a mandatory risk management activity for many financial institutions around the world. In addition, large fluctuations in financial markets can also have a broad impact on the economy. For example, the financial crisis of 2008 caused huge volatility in financial markets, which had a negative impact on the world economy. Therefore, the study of model construction, theoretical analysis, numerical solution and application of volatility analysis and forecasting has become an important issue in the field of financial time series analysis.

2.2.1. ARCH and GARCH Type Models

A conditional heteroskedasticity model is an econometric model that models asset returns and their volatility.

Engle (1982) introduced the autoregressive conditional heteroskedasticity (ARCH) model, a model that is still considered a valid tool for measuring volatility in time-varying financial markets. On this basis, a large number of studies have been conducted based on ARCH models for estimating financial time series volatility.

Engle et al. (1987) proposed the ARCH-M model by adding a factor of condi-

tional variance to the conditional mean equation. In 1991, Robinson (1991) proposed a linear ARCH model in the derivation of the heteroskedasticity test.

Bera et al. (1992) proposed the Augmented ARCH model by adding the interaction between past perturbations of different orders to the conditional variance equation of the above ARCH model. In the same year, Christian and Monfort (1992) expressed the conditional variance as a summation form of a column of step functions, thus proposing the QTARCH model. Guégan and Diebolt (1994) proposed the Beta ARCH model for volatility asymmetry.

Donaldson and Kamstra (1997) gave the ANN ARCH model by incorporating the logistic function, which is widely used in the field of neural networks, into the ARCH model.

Later, Han and Park (2008) introduced the ARCH-NNH model. This model is based on the ARCH model by adding a non-linear function of the unit root process and doing so in order to investigate the effect of slow decaying variables on volatility. Li et al. (2016) constructed the TDAR model by taking into account the conditional heteroskedasticity on top of the TAR model.

Bollerslev (1986) proposed the generalised autoregressive conditional heteroskedasticity model, also known as the GARCH model, in 1986. This model involves adding the autoregressive term of the prior moment conditional variance to the conditional variance equation of the ARCH model. Since financial return series tend to exhibit weak correlation while squared return series are in most cases highly correlated. Therefore, similar to ARCH models, a large number of scholars use GARCH-type models to analyse and forecast volatility.

Similar to the GARCH model, Taylor (1986) and William Schwert (1989) construct regression equations for conditional standard deviations to propose the TS-GARCH model. Taking logarithms of the variables in the conditional variance equation, Dieobold (1986), Pantula (1986) and Milhøj (1987) proposed log-GARCH models.

Since in many practical applications the standard residuals obey a thicktailed distribution different from the normal distribution, Bollerslev (1987) added the standard Student's t distribution to the GARCH model and used this to fit the standard residuals. Making the conditional variance dependent on a non-linear transformation of the squared perturbation term, Friedman et al. (1989) proposed the Modified ARCH model. Taking the nth order term of the absolute value of the perturbation as the independent variable and the nth order of the conditional standard deviation as the dependent variable, Higgins and Bera (1992) proposed a non-linear GARCH model.

Sentana (1995) introduced the generalised quadratic ARCH model, which portrays the effect of the interaction of different orders of perturbations on the conditional variance by adding the quadratic term of the perturbation to the conditional variance equation.

Bollerslev and Ghysels (1996) proposed the cyclical GARCH model taking into account the factor of periodicity.

To study the correlation of long-term volatility, Lee and Engle (1993) proposed the Component GARCH model in 1999. In 2001, Nowicka-Zagrajek et al. (2001) replaced the constant term truncation parameter in the GARCH model with the summation of a sequence of independent identically distributed random variables, thus proposing the Randomized GARCH model.

In response to the jump in volatility, Maheu and McCurdy (2004) proposed the GARJI model based on the GARCH model. In 2006, Alexander and Lazar (2006) proposed the NM-GARCH model based on a mixed distribution where the conditional standard deviation follows a normal distribution.

In 2015, Ahmed (2016) proposed a conditional heteroskedasticity binary choice model for macro-financial time series. Ahmad and Francq (2016) introduced a Poisson fitted maximum likelihood estimation procedure in an INGARCH model for integervalued time series.

2.2.2. Asymmetric Effects

A key feature for financial data is the asymmetric effect of positive and negative returns on volatility. With the discovery of this asymmetry, a large body of research has begun to introduce asymmetric structures into models to interpret this financial phenomenon.

In 1990, Engle (1990) constructed the AGARCH model by introducing a perturbed primary term into the GARCH model to account for the asymmetry of volatility. In 1991, Nelson (1991) introduced the EGARCH model, which defined the conditional variance in logarithmic form to more strongly reflect the asymmetry of volatility, while avoiding the non-negativity constraint that some model parameters must satisfy.

In 1993, Ding et al. (1993) introduced the APGARCH model, in which they added the nth order term of the perturbation to the non-linear GARCH model to measure volatility asymmetry. Glosten et al. (1993) added the product of the squared perturbation term and the schematic function of the perturbation term to the GARCH model to characterise volatility asymmetry, also known as the GJR model.

In 1994, Zakoian (1994) used a schematic function to describe the conditional standard deviation and introduced a dummy variable d_{t-1} into the conditional variance, distinguishing the effects of positive and negative shocks on volatility through different values of the dummy variable.

In 1996, Fornari and Mele (1996) proposed the VSGARCH model based on the GJR model, making volatility asymmetry dependent not only on past perturbations but also on past conditional variances. In the same year, Li and Li (1996) proposed a two-layer threshold ARCH model that gives a threshold structure to the parameters in the conditional variance equation.

In 1998, Gonzalez-Rivera (1998) introduced a smooth transformation function into the GARCH model so that the conditional variance is related not only to the positivity or negativity of the disturbance but also to the magnitude of the disturbance, i.e. the Smooth Transition GARCH model.

In 2006, Caporin and McAleer (2006) introduced multiple threshold institutions on top of the GJR model, thus proposing the DAGARCH model. In 2017, Takaishi (2017) introduced rational functions into the GARCH model to interpret volatility asymmetry, i.e. the Rational GARCH model.

The above people use data from the financial sector itself, such as stock data, futures data, fund data, and data reflecting economic indices. Most of these data are economic in nature, and do not cover public sentiment, or news sentiment. So this paper adds these variables to the study, such as news sentiment, internet user sentiment, and tweet sentiment from Twitter.

2.3. Investor Sentiment

2.3.1. Traditional Investor Sentiment

The specific definition of investor sentiment is divided into two main areas: firstly, "limited rationality", irrational emotional judgements arising from the psychological utility of investors; and secondly, "cognitive bias", investors' expectations of future asset price movements and behavioural bias.

Barberis et al. (2005) was the first to suggest that when investors adopt irrational emotions to improvise subjective choices and determine trades, they violate the trading laws of expected utility theory, i.e. for choices under investor sentiment, which very much emphasizes the importance of psychology to the study of financial markets, market trading and the choice of financial products, where psychological behaviour ultimately affects actual decisions. Baker and Wurgler (2006) point out that investor sentiment for the study of investors' future expectations of the market and the adoption of behaviours such as confidence, representativeness or conservatism caused by bias benefits, summarised as a belief that investors do not rely on financial statement cash flows and trading risk and trade.

2.3.2. Metrics of Investor Sentiment

To explore the relationship between investor sentiment and the stock market from the perspective of behavioural finance, the establishment of investor sentiment indicators is of paramount importance. Firstly, the sentiment indicators reflecting the sentiment of investors in the market are extracted from market transactions, and a comprehensive index of investor sentiment is synthesised using relevant methods. So far, investor sentiment indicators can be broadly categorised into 1) direct indicators obtained from surveys, 2) indirect indicators from data statistics and 3) internet opinion indicators from text mining.

Direct indicators generally take the form of questionnaires, which are aggregated by counting participants' expectations and perceptions of the future, and can directly reflect investors' most realistic sentiments and perceptions of the market. Fisher and Staman (2003) found that the American Association of Individual Investors Index can be used as a predictor of the inverse of the S & P500 future returns.

A single investor sentiment indicator is one that uses a single proxy variable

when measuring investor sentiment. Liu (2015) explored the liquidity indicator to uncover its relationship with investor sentiment and used Granger causality tests to demonstrate that there is a causal relationship between investor sentiment and market liquidity, which can be used as an objective indicator. Kumar and Lee (2006) used the proportion of zero shares bought and sold as an indicator of small and medium-sized investors' sentiment in the market and tested it using cross-section and obtained that the indicator has a greater impact on the returns of small-cap stocks, value stocks, low-priced stocks and stocks with low institutional participation.

A composite investor sentiment index refers to the selection of multiple proxies, the extraction of factors common to multiple proxies, and the use of principal component analysis to reduce dimensionality.

Baker and Wurgler (2006) were the first to propose a BW index, which was constructed by selecting six proxies to exclude macroeconomic influences and proved to be more explanatory than a single indicator. Brown and Cliff (2005) used investor sentiment and long-term asset valuations and returns constructed using fund discounts to conclude that large market capitalisation and low book value portfolios can cause investors to push prices above their intrinsic value when they are overly optimistic, and that sentiment and valuation errors are positively correlated and past returns are predictive of future returns. Raissi and Missaoui (2015) use principal component analysis, a measure of investor sentiment, to investigate the correlation between investor sentiment and stock market liquidity and returns, respectively, in their study of stock markets. Stambaugh et al. (2012) obtained good results by examining the excess returns, high and low book-to-market P/E ratios, and small and large company earnings differentials on returns for investor sentiment construction in the stock market.

2.3.3. Internet Based Investor Sentiment

As a branch of data mining, text mining has developed rapidly in recent years. The explosive development of the Internet has led to the release of a large amount of original data on web platforms every day, and merely obtaining structured data information is no longer enough to conduct a comprehensive and accurate study of the facts. Text mining technology can achieve effective extraction of unstructured web information, through feature extraction, text classification, text clustering, semantic analysis, etc. to achieve text database information extraction. With the popularity of the Internet, the analysis of socio-economic behaviour based on web search data has gradually become a new hot topic, and relevant text mining and semantic analysis techniques have also made great progress.

In recent years, text mining sentiment indicators based on text analysis of online media content have been increasingly used in behavioural finance research.

Tetlock (2007) used media to linguistically characterise positive and negative words in corporate news reports, concluding that negative words have greater

returns and predictability and that investors capture media reports to inform investment decisions.

Choi and Varian (2012) used Google engine data to make short-term economic forecasts of behaviour including car sales, household sales, and retail sales, and concluded that search engine data can, to some extent, reflect investor sentiment in order to predict short-term economic values.

Nassirtoussi et al. (2014) point out that text mining techniques combining interdisciplinary areas such as natural language processing, behavioural economics and artificial intelligence have become an emerging approach to market forecasting techniques, and the article makes a significant contribution to research on sentiment interpretation and market forecasting.

Da et al. (2015) used millions of household daily internet searches to explain market-level sentiment, and search terms such as household question queries as a new indicator of investor sentiment, finding that this indicator was effective in predicting short-term return reversals and increased volatility phenomena.

You et al. (2017) used a web crawler approach to study Twitter content and obtained that the investor sentiment indicator also had significant predictive power for country market returns.

Those above, on the other hand, do not adequately consider various types of economic indices when considering investor sentiment, but only a single type of economic indicator, whereas this paper will consider the share prices of five vaccine producing companies and correlate tweet sentiment with company share prices.

3. Data

The data for this paper was selected from 31 December 2019 to 9 July 2021. This time period was chosen because the Covid-19 pneumonia epidemic officially began to emerge globally in early 2020, with an outbreak in Wuhan, China in February 2020, which in turn became contagious worldwide. On 19 July 2021, the official decontrol of the disease is announced in the UK and 80% of the UK population has been vaccinated against Covid-19. The entire year and a half therefore covers the impact of the entire phase of virus transmission and vaccine development. Table 1 presents descriptive statistics for the data required for this paper.

3.1. Pharma Stock Data

With the outbreak of the epidemic, companies developing vaccines have become the focus of the stock market. The stock data of pharmaceutical companies is therefore the focus of our consideration. In the latest vaccine development process, five companies—Pfizer and AstraZeneca in the UK, Moderna and Novavax in the USA and BioNTech in Germany, produce the world's leading vaccines. We have therefore selected the stock prices of these five companies for our research. Table 2 shows summary statistics for the closing prices of the shares of

Table 1. Summary of statistics data.

Variables	Description	Mean	Std. Dev
Stock log-Returns			
R_{BNTX}	Daily BioNTech stock's log-returns	0.0048	0.0688
R_{MRNA}	Daily Moderna stock's log-returns	0.0065	0.0563
R_{PFE}	Daily Pfizer stock's log-returns	0.0002	0.0193
R_{AZN}	Daily AstraZeneca stock's log-returns	0.0005	0.0201
R_{NVAX}	Daily Novavax stock's log-returns	0.0100	0.0905
Twitter Sentiment			
POT	Daily average polarity point of all tweets	0.0477	0.0334
PPOT	Daily average polarity point of positive tweets	0.2510	0.0238
NPOT	Daily average polarity point of negative tweets	-0.2201	0.0257

Table 2. Closing prices of stocks.

Stock	Min	*1ˢᵗ*	Median	Mean	*3ʳᵈ*	Max
BioNTech (BNTX)	28.55	50.78	85.03	93.44	112.41	241.49
Moderna (MRNA)	17.78	57.40	72.97	96.39	147.06	235.11
Pfizer (PFE)	27.03	34.39	35.87	35.78	37.20	42.56
AstroZeneca (AZN)	37.79	49.93	52.56	52.20	55.06	61.10
Novovax (NVAX)	3.93	40.46	111.08	109.44	168.10	319.93

the five companies between 31 December 2019 and 9 July 2021.

For most investors, stock returns are independent of the size of the investment and are a good indicator of the stock's investment opportunities, while yield series are easier to handle due to their better statistical properties than stock price series, so most financial studies have focused on stock yield series rather than their price series.

In this paper, the stock log-return is used as a characteristic to study stock volatility. The daily closing price of a stock is calculated to obtain the stock's log-return $R_{t,i}$, with the following formula.

$$R_{t,i} = \ln\left(P_{t,i} / P_{t-1,i}\right), \tag{1}$$

where $P_{t,i}$ is the closing price of the stock at time t, and $i \in \{\text{BNTX, MRNA, PFE, AZN, NVAX}\}$ represents five pharmaceutical companies.

In this paper, the daily closing prices and daily log-returns of the five selected stocks are plotted in Figures 1-5 below. Figures 1-5 are from https://finance.yahoo.com/.

3.2. Twitter Sentiment Data

This article captures the Twitter data on the Internet. The time period of the

Figure 1. Closing price and log-returns of BNTX. (a) Closing price; (b) Log-returns.

Figure 2. Closing price and log-returns of MRNA. (a) Closing price; (b) Log-returns.

Figure 3. Closing price and log-returns of PFE. (a) Closing price; (b) Log-returns.
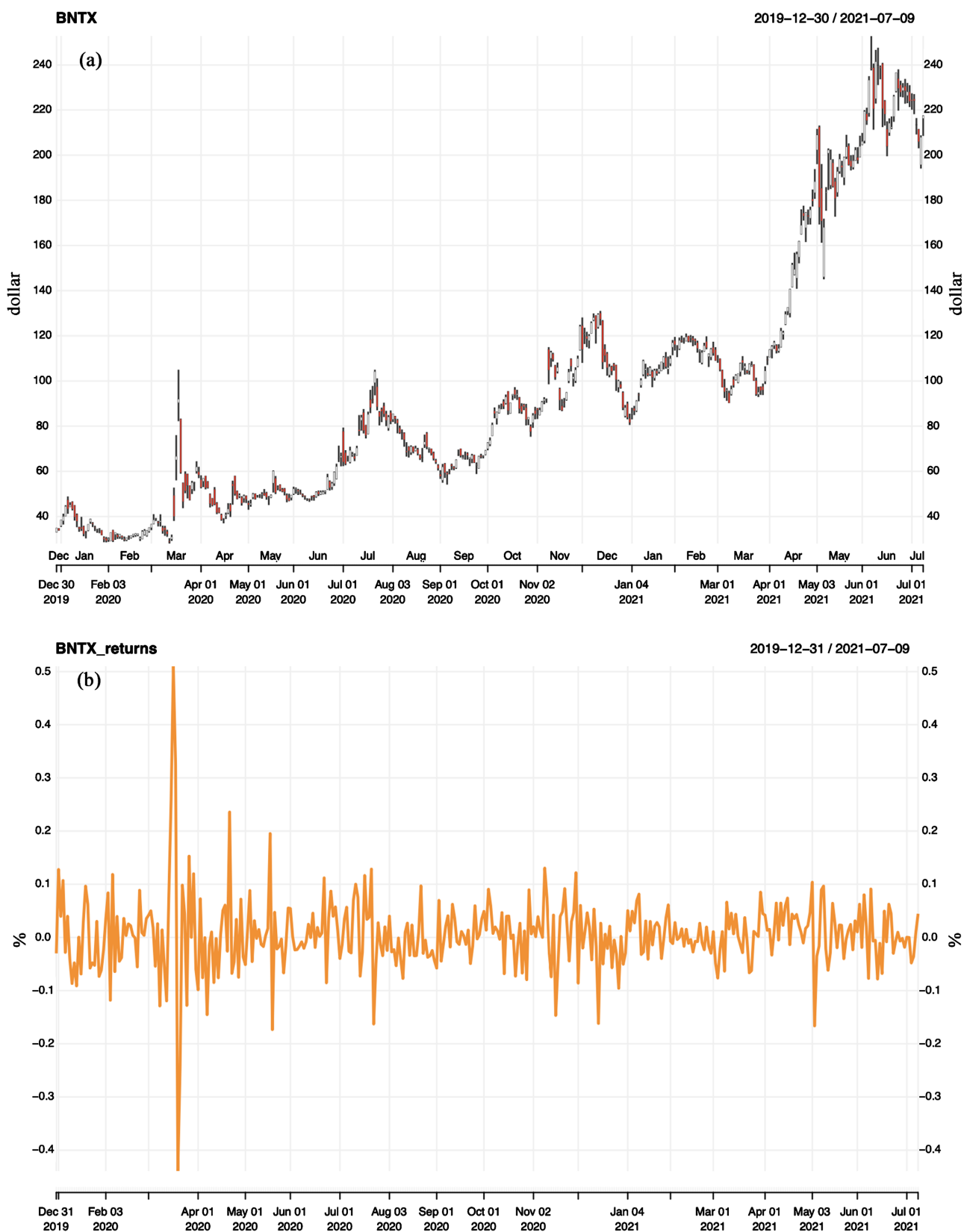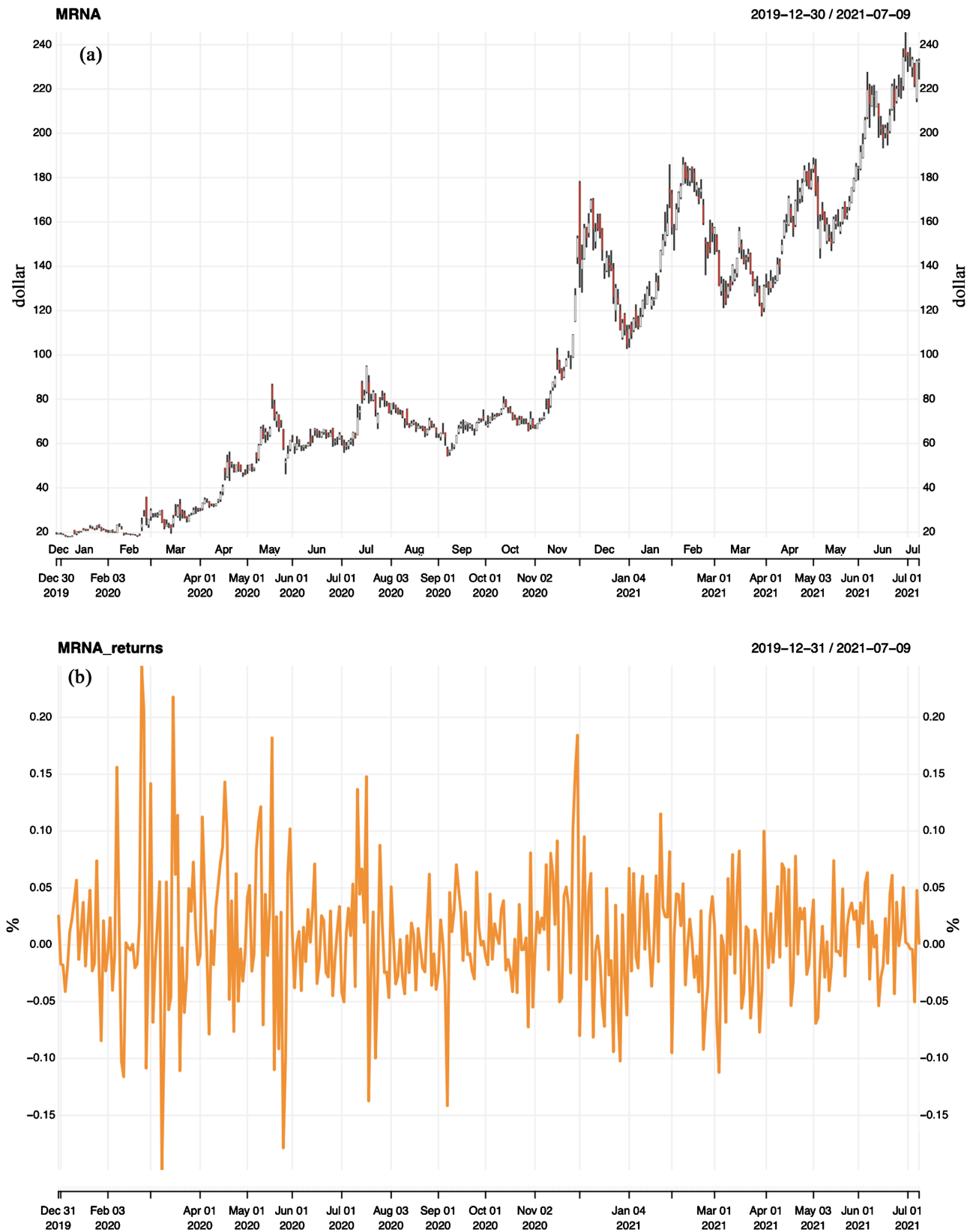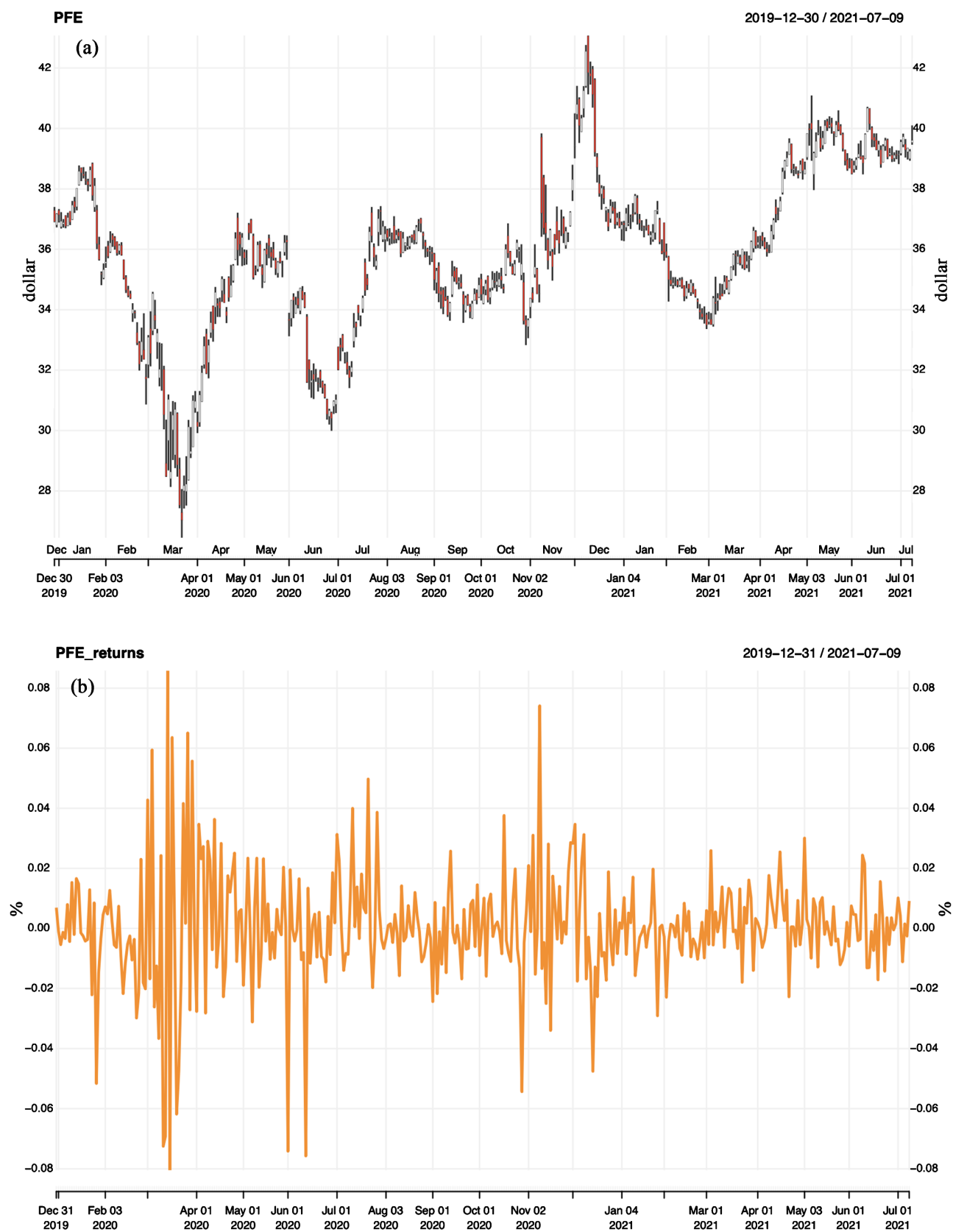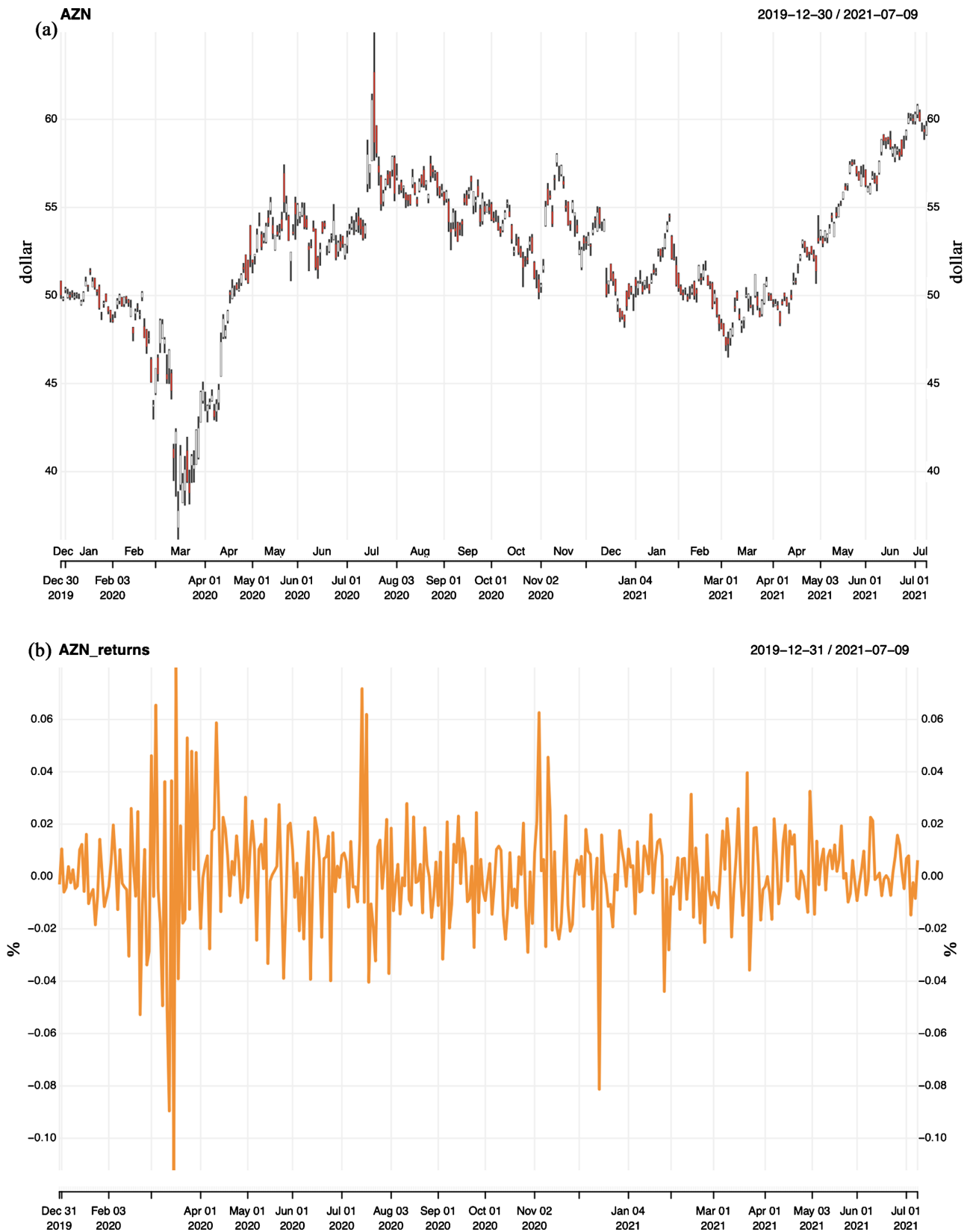
Figure 4. Closing price and log-returns of AZN. (a) Closing price; (b) Log-returns.

Figure 5. Closing price and log-returns of NVAX. (a) Closing price; (b) Log-returns.

Twitter data is consistent with the previously selected stock time, which is from December 31, 2019 to July 9, 2021. The keywords in the Twitter data must be related to Covid-19 and vaccines.

Therefore, two lists are set up, the words of the first list are related to covid-19, and the words of the second list are related to vaccines. When fetching tweets, the tweets must match the keywords in the list. That is, at least one word in each list is included in the crawled tweet:

- List 1: Covid, Covid-19, covid, coronavirus, Covid-19, COVID.
- List 2: Vaccinate, Jab, jab, Shot, shot, Vac, Vaccine, Vaccination, vaccinating, vaccinations, vaccines.

This article uses the Twint package in Python to capture Twitter data. Twint is a Twitter crawling tool written in Python that allows you to crawl tweets from Twitter configuration files without using Twitter's API. It can be used anonymously, no Twitter registration is required and there is no rate limit.

After several hours of crawling, we extracted a total of 2,728,128 tweets. After data cleaning, duplicate tweets and non-English texts were deleted. Because in the following analysis, the sentiment analysis algorithm is limited by language, so it must be limited to the same language, so only the English text is retained. In addition, the English version of Twitter also occupies the main part, and the effect of the analysis is also very good. The total number of tweets obtained after the final data cleaning is 1,795,552.

Usually, tags are used to extract tweets. However, we searched for keywords in the main text to extract tweets. This approach allowed us to still crawl most of the tweets discussing the Covid-19. Therefore it is effective.

Figure 6 shows the average daily number of tweets from the emergence of Covid-19 and its evolution to a global pandemic until July 9, 2021, which mentioned viruses and vaccines in the tweets we crawled. The graph shows that there was almost no discussion on the topic of viruses and vaccines before the virus
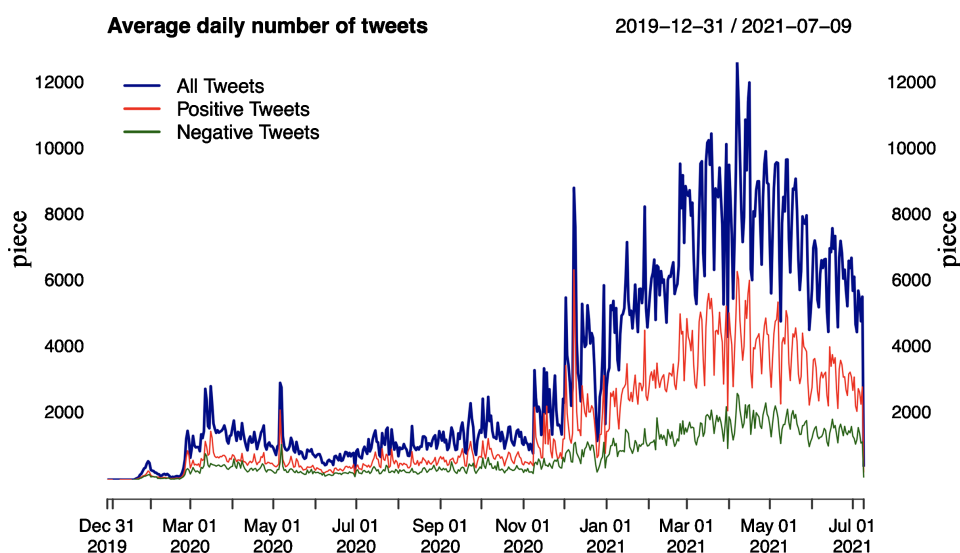

Figure 6. Daily number of tweets mentioning covid vaccine about English style.

began to spread. When the virus began to spread globally in March 2020, tweets discussing viruses and vaccines began to increase gradually and reached the first small peak. After that, the number of discussions went up and reached the second peak in November 2020, when the five leading companies in vaccine research and development announced the efficacy of their vaccines. Therefore, more and more people discuss vaccines. The number of tweets continued to increase. By April 2021, the number of tweets reached its peak, because at this time both the Europe and the United States began to be able to vaccinate in large quantities, and the public began to vaccinate on a large scale. Moreover, the delta variant virus has also begun to spread globally, making tweets discussing viruses and vaccines high. After that, the number of tweets declined, indicating that the delta variant virus began to no longer show explosive growth after the population was vaccinated, and the atmosphere of discussion declined, but the heat remained high.

Twitter Sentiment Analysis

Next, begin to analyze the Twitter data. We use the Internationally popular TextBlob library in Python. This library can determine the subjectivity and polarity of each tweet in the crawled data and assign it a certain value. The operating method of the library is to use Pattern Analyzer to recognize parts of speech for natural language processing, and then construct a pattern graph, and finally recognize the sentiment of the text. The specific principle is shown in Figure 7.

In terms of sentiment, we start by identifying the point of view of each tweet. This is done by assigning and classifying polarities. Polarity is the sentiment revealed in the text and there are three types of sentiment: positive, negative, and neutral. The score for sentiment fluctuates from −1 to 1. −1 means completely negative, 1 means completely positive, and 0 means neutral. A score of 0 to text can be completely neutral or can result from a mixture of negative and positive.
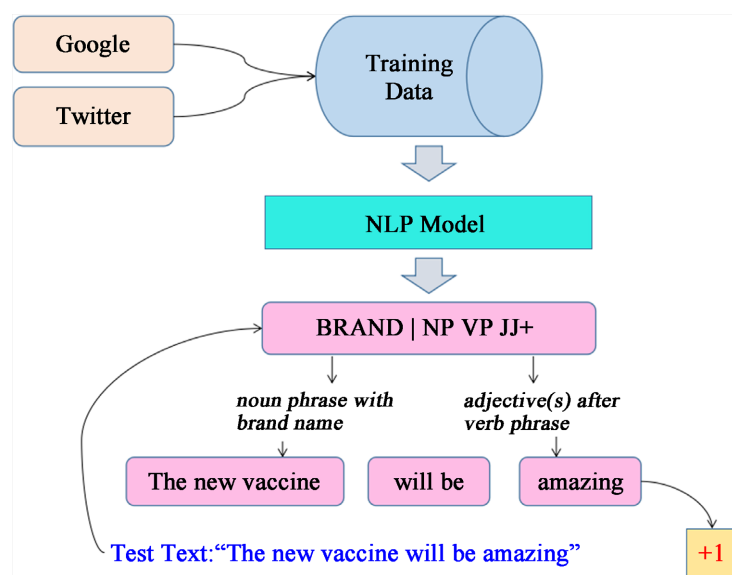

Figure 7. TextBlob Pattern.

When calculating text polarity, speech labels are important; each word has a corresponding type and meaning in the thesaurus, and is later assigned a corresponding polarity score. And when calculating the sentiment of individual words, the thesaurus takes the average value as the final output.

And to calculate the sentiment of a set of texts, we will consider the average sentiment score of the phrases in the text. For example, for a tweet, we multiply the sentiment of each word in the sentence to determine the total sentence score, but also consider negation and intensity. When a negation was encountered, the phrase was multiplied by −0.5 to deal with the negation. Ignore words that do not appear in the training set or are one-word words. If the utterance has a modifier, assign an intensity value to it and multiply or divide by the presence of the negative utterance.

When it comes to natural language processing, there are some problems that are difficult to deal with. Irony, for example, this is common both in life and in tweets, and is difficult to quantify. Yet thanks to the exclamation mark, our ability to capture them is greatly enhanced.

Moreover, in sentiment analysis, polarity and sentiment are interchangeable because sentiment is reflected by relying on polarity.

4. TGARCH-M Model

Most of the theoretical premises of the stock market are based on the fact that 1) the perturbation terms of returns are independent of each other and 2) the variance is constant. But as financial theory has evolved and research has intensified, we have found that stock market prices fluctuate from time to time and that this volatility tends to have an aggregated nature, i.e. large fluctuations are clustered together in frequent and violent swings. Smaller fluctuations are often followed by smaller fluctuations, which are moderated, meaning that fluctuations are time-varying and the variance of fluctuations is not constant, but also changing.

As a result, time series models that assume constant variance are no longer suitable for describing stock market volatility, and many experts and scholars began to look for a new approach to the study of stock market volatility. Many scholars have since discovered that GARCH-type models, i.e. conditional heteroskedasticity-type models, can better describe the time-varying characteristics of volatility, and so various GARCH-type models and their related extensions have emerged, which are described below.

4.1. GARCH-Type Models

GARCH-type models are derived from ARCH models with varying degrees of refinement, and the various settings in the models have their own economic implications. This paper therefore introduces each model from the ARCH model to the TGARCH-M model, and explains the reasons for the choice of model and the economic implications of each setting.

4.1.1. ARCH Model

The ARCH model was discovered in 1982 by Engle (1982) in his analysis of macro data and its main core is that the variance σ^2 of the perturbation term ε at moment t is influenced by the magnitude of the squared error at moment $t-1$, which is dependent on ε_{t-1}^2.

The core of the model is to set the process of generating the perturbation term ε_t as:

$$\varepsilon_t = v_t \sqrt{\alpha_0 + \sum_{i=1}^{n} \alpha_i \varepsilon_{t-i}}, \tag{2}$$

In Equation (2), $\alpha_0 < 0$, $0 < \alpha_i < 1$, v_t is white noise and is independent of ε_t. The conditional variance of ε_t is:

$$\sigma_t = Var\left(\varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_{t-n}\right) = \alpha_0 + \sum_{i=1}^{n} \alpha_i E\left(\varepsilon_{t-i}^2\right), \tag{3}$$

Equation (3) is the defining equation for ARCH (q) model, where the variance in the current period depends on the linear combination of the variances of the previous periods. In addition, the magnitude of $Var(b)$ depends not only on the variance of the previous periods but also on the sensitivity to the previous variance α_i. When the squared residuals in the regression results have a non-zero autocorrelation coefficient, they can be modelled by the ARCH model. The ARCH (q) model is set up as:

$$\begin{aligned} y_t &= f\left(x_t, x_{t-1}, \cdots\right) + \varepsilon_t \\ \varepsilon_t \mid I_{t-1} &\sim N\left(0, \sigma_t^2\right) \\ \varepsilon_t &= v_t \sqrt{\alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-q}} = v_t \sqrt{\sigma_t^2} = v_t \sqrt{h_t}, \end{aligned} \tag{4}$$

where y_t denotes the explained variable at moment t, which is explained by the explanatory variable $X_t = \left(x_1, x_2, \cdots, x_t\right)$. ε_t is the perturbation term at moment t, I_t is the set of information at moment t, σ_t^2 is the conditional variance of ε_t, ω is a constant, and ε_{t-i}^2 is the squared residual of the lag. $\omega > 0$, $\alpha_i > 0$, $\sum \alpha_i < 1$. This ensures that the ARCH process is smooth. At this point, ε_t obeys the ARCH (q) process. And the conditional variance of ε_t is known from the expression of the conditional variance σ_t^2 as a linear combination of $\left(\varepsilon_{t-1}^2, \cdots, \varepsilon_{t-q}^2\right)$ of a linear combination. When ε_{t-1} is so big, σ_t can become big, indicating that future market volatility will be positively influenced by regression disturbance terms from the past. The magnitude of the value of q can decide how long a particular jump in the random variable continues to affect. So the ARCH model can reflect the aggregation of volatility in the stock market.

Once the ARCH model was proposed, it became one of the most important methods for studying heteroskedasticity in econometrics. But there are also certain drawbacks. For example, in the ARCH (q) model, ε_t is often set to obey a normal distribution, but in practice most financial time series are characterised by spikes and thick tails, and the normal distribution is inaccurate. The second is

that ε_t^2 in the ARCH (q) model is considered to be an even function of the information ε_t. This conclusion is unreasonable because the magnitude of ε_t^2 does not depend only on the absolute value of ε_{t-1}, but is also affected by its positive and negative effects, which is inconsistent with the leverage effect of financial problems. But in fact, future volatility and current returns are always correlated negatively in the stock market, and the value of the ARCH model's conditional variance depends on the value of the new data, and has nothing to do with the trend reflected in the new information with it, and does not make full use of the information provided by the new information. The third is that in practice, the order q needs to be large in order to achieve good results, which inevitably increases the computational effort.

The ARCH model has the advantage of being able to accurately adjust for changes in financial time series volatility and is broadly used in practical studies of time series financial analysis, thus allowing investors to more accurately capture risk.

The disadvantage of ARCH models is that they generally assume that positive and negative shocks have the same response to volatility. ARCH models are very tightly parameterised and offer no new insights into the sources of time series variation, but simply provide a way to reflect the state of the variance.

4.1.2. GARCH Model

In 1986, Bollerslev (1986) extended and improved the ARCH model in order to more accurately characterize the distribution of the tails of the time series, proposing the GARCH model, thus making the latter's lagged results more flexible. The GARCH model is a model that introduces its own lagged values in the determination of the conditional variance σ_t^2 of the current period of ε_t, the GARCH (p, q) model is set up as:

Mean equation:

$$
\begin{aligned}
y_t &= f\left(x_t, x_{t-1}, \cdots\right) + \varepsilon_t \\
\varepsilon_t \mid I_{t-1} &\sim N\left(0, \sigma_t^2\right) \\
\varepsilon_t &= v_t \sqrt{h_t}
\end{aligned}
\tag{5}
$$

Variance equation:

$$
h_t = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \beta_i h_{t-i} + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2,
\tag{6}
$$

where σ_t^2 is the conditional variance based on the past correlation information, and p is the order of the autoregressive GARCH term, then q is the order of the ARCH term. And we must ensure that the conditional variance is positive, $\alpha_0 > 0$, $\alpha_j > 0$, $\beta_i > 0$, $\sum \alpha_j + \sum \beta_i < 1$.

The magnitude of the model's coefficient $\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j$ reflects the persistence of the series' volatility, i.e. the size characteristics of the series' volatility at past moments are inherited at the current moment. If the value of $\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j$ is closer to 1, the more is inherited and the greater the vola-

tility of the whole series. When this value is less than 1, it means that the shock at a given moment will fade away, but when this value is greater than 1, it means that the impact of this shock will not fade away and is spread.

Although the GARCH (p, q) model has a great deal of applicability over the ARCH model, it has the following shortcomings in terms of applied assets:

1) This is not well explained by the GARCH model, which assumes that the variance varies with the square of the residuals, when the desired return on an investment in equities is negatively correlated with the volatility of the change in returns. Therefore, the change in volatility is not affected by the positive or negative variance of the residuals. The GARCH model does not describe this asymmetric relationship.

2) The GARCH model supposes that the coefficients in the volatility function are all larger than zero in order to satisfy the non-negative condition.

3) When using the GARCH model, the variance is determined by taking into account only the magnitude, but not the sign. In other words, the model does not distinguish between negative and positive shocks. This problem can be solved by using a TARCH model, which means that another variable can be introduced so that the disturbance term does not change to the same extent when it is negative and positive, which makes the description more accurate and reasonable.

4.1.3. GARCH-M Model

The conditional mean of the return on some financial assets is affected by their volatility, known as the risk premium.

In time series data, sometimes there is a relationship between the variance and the mean of the series. For example, there is a positive correlation between stock prices and risk represented by variance. We think that the conditional variance may be applied as a metric of risk that changes over time, thus linking return and risk. Therefore, in the mean value equation, variance is one of the important explanatory variables and the omission of a variance term that is correlated with an explanatory variable can create endogeneity problems. The GARCH-M model is obtained by adding the variance term to the mean model and the GARCH-M model is set up as:

Mean equation:

$$
\begin{aligned}
&y_t = f(x_t, x_{t-1}, \cdots) + \tau h_t + \varepsilon_t \\
&\varepsilon_t \mid I_{t-1} \sim N(0, \sigma_t^2) \\
&\varepsilon_t = v_t \sqrt{h_t}
\end{aligned}
\tag{7}
$$

Variance equation:

$$
h_t = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \beta_i h_{t-i} + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2,
\tag{8}
$$

The model aims to explain the returns on financial assets by increasing σ_t because every investor expects asset returns to be closely linked to risk, and the conditional variance σ_t^2 reflects the magnitude of the expected risk. The para-

meter τ is called the risk premium parameter, and if τ is positive then returns are positively correlated with volatility.

4.1.4. TGARCH Model

The TGARCH model, introduced by Zakoian (1994), Glosten et al. (1993), and Zakoian (1994), is a better simulation of the "leverage effect" in financial markets. For example, stock prices in the stock market are influenced by market news. The exogenous variables can be added to the variance equation and the TGARCH model with the exogenous variables is set up as follows:

Mean equation:

$$
\begin{aligned}
y_t &= f\left(x_t, x_{t-1}, \cdots\right) + \varepsilon_t \\
\varepsilon_t \mid I_{t-1} &\sim N\left(0, \sigma_t^2\right) \\
\varepsilon_t &= v_t \sqrt{h_t}
\end{aligned}
\tag{9}
$$

Variance equation:

$$
h_t = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \beta_i h_{t-i} + \sum_{j=1}^{q}\left(\alpha_j \varepsilon_{t-j}^2 + \gamma_j D_{t-j} \varepsilon_{t-j}^2\right),
\tag{10}
$$

where $\sum_{j=1}^{q} \gamma_j D_{t-j} \varepsilon_{t-j}^2$ is the TGARCH term.

When $\sum_{j=1}^{q} \gamma_j = 0$, there is an asymmetric effect. And if $\sum_{j=1}^{q} \gamma_j \neq 0$, there is no asymmetric effect.

And we set D_{t-j} as a dummy variable. When $\varepsilon_{t-j} < 0$, $D_{t-j} = 1$. This means that "bad news" has occured. When $\varepsilon_{t-j} > 0$, $D_{t-j} = 0$. It means that "good news" has occured. At this moment, the conditional variance is not additionally affected by the news.

When $\sum_{j=1}^{q} \gamma_j > 0$, the asymmetric effect exists and is positive. This means that the emergence of bad news has a greater negative impact on the stock than the emergence of good news has a positive impact on the stock. This is reflected in volatility, where good news has an α shock and bad news has an $(\alpha + \gamma)$ shock.

When $\sum_{j=1}^{q} \gamma_j < 0$, the asymmetric effect exists and is negative. In other words, good news has a larger shock. The occurrence of good news increases the volatility of the stock market, while the occurrence of bad news reduces volatility.

4.2. TGARCH-M Model Specification

Combining the GARCH-M model with the TGARCH model, and considering the relationship between conditional variance and mean variance, and considering the effect of exogenous variables on the variance equation, constitutes the TGARCH-M model.

TGARCH-M model is set up as:

Mean equation:

$$
\begin{aligned}
y_t &= f\left(x_t, x_{t-1}, \cdots\right) + \tau h_t + \varepsilon_t \\
\varepsilon_t \mid I_{t-1} &\sim N\left(0, \sigma_t^2\right) \\
\varepsilon_t &= v_t \sqrt{h_t}
\end{aligned}
\tag{11}
$$

Variance equation:

$$h_t = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \beta_i h_{t-i} + \sum_{j=1}^{q} \left(\alpha_j \varepsilon_{t-j}^2 + \gamma_j D_{t-j} \varepsilon_{t-j}^2 \right), \tag{12}$$

5. Twitter Sentiment Index

Antweiler and Frank (2004) proposed a method to construct a bullish indicator based on the classification of stock forum posts. Therefore the following Twitter Sentiment indicators are constructed in this paper.

$$B_t = \frac{M_t^{pos} - M_t^{neg}}{M_t^{pos} + M_t^{neg}}, \tag{13}$$

where $M_t = \sum \omega_i x_i^c \left(i \in D(t) \right)$ denotes the sum of the weighted number of messages of type $c \in \{ pos, neu, neg \}$ over a period of time D_t. Where pos represents positive sentiment, neg represents negative sentiment and neu represents neutral sentiment. x_i^c is the indicator variable that takes on the value of 1 if message i belongs to a certain type c and 0 otherwise. In particular, when the weights are all equal to 1, M_t^c is equal to the total number of messages of type c over the time period $D(t)$. The Twitter sentiment indicator B_t, which lies between −1 and 1, expresses the relative bullishness of Twitter users, and the indicator is independent of the total number of tweets. In addition, Antweiler and Frank define another indicator:

$$B_t^* = \ln \left[\frac{1 + M_t^{pos}}{1 + M_t^{neg}} \right], \tag{14}$$

Then they think that $B_t^* \approx B_t \ln \left(1 + \left(M_t^{pos} + M_t^{neg} \right) \right)$. B_t^* takes into account not only the relative degree of positivity, but also the number of posts expressing positive and negative tweets. The B_t^* indicator is also shown to be superior in their study.

Although B_t^* considers the number of tweets, it only considers the number of tweets expressing positive and negative sentiment and does not consider the number of neutral tweets. This paper argues that the level of user attention is also an expression of user sentiment, so even if users express neutral expectations, this information is still valuable. In view of this, this paper proposes a fusion of B_t and the total number of tweets, to express the Twitter sentiment index $B_t^{twitter}$ of user attention.

$$B_t^{twitter} = B_t \ln \left(1 + M_t \right), \tag{15}$$

where $M_t = M_t^{pos} + M_t^{neg} + M_t^{neu}$.

In addition, to indicate positive and negative sentiments, this paper makes

$$dB_t^{twitter} = B_t^{twitter} - B_{t-1}^{twitter}, \tag{16}$$

When $dB_t^{twitter} > 0$, it means that the user turns optimistic or positive in period t. When $dB_t^{twitter} < 0$, it indicates that the user turns pessimistic or more negative in period t.

The number of tweets is so large that it may make the range of variation of the Twitter sentiment index larger. To solve this problem, the $dB_t^{twitter}$ is normalised so that the range of variation is between −1 and 1.

$$dB_{t,scale}^{twitter} = 2 \frac{dB_t^{twitter} - \min\left(dB_t^{twitter}\right)}{\max\left(dB_t^{twitter}\right) - \min\left(dB_t^{twitter}\right)} - 1, \tag{17}$$

6. Correlation Tests for Empirical Indicators

Before carrying out the relevant empirical analysis, we usually perform relevant tests on the financial time series, such as correlation, stationarity, conditional heteroskedasticity, etc., in order to prepare for the subsequent empirical analysis.

6.1. Descriptive Statistical Analysis

Table 3 presents a basic descriptive analysis of the log returns of the five pharmaceutical companies. It can be seen that both BioNTech and Novavax have skewnesses greater than 0 and kurtosis greater than 3, indicating that both show a spiky right skew pattern. The skewness of Moderna is greater than 0, but the kurtosis is less than 3, indicating a right-skewed pattern. For Pfizer and AstraZeneca, the skewness was less than 0 and the kurtosis was greater than 3, indicating a spiky left skewed pattern. The JB statistic for all five companies rejects the normality of the log returns.

6.2. Sequence Smoothness Test

The pseudo-regression problem is a key consideration in the modelling and analysis process and must be avoided at all costs. Pseudo-regressions should be eliminated so that the time series can be made smooth. There are many ways to test for series smoothness, including the ADF unit root test, the ERS test, the PP test, and the NP test. Here we use the common ADF test.

According to the results of the unit root test in Table 4, the p-values for the log returns of stocks and the Twitter sentiment indicator are less than 0.05, so the null hypothesis is rejected. All series data do not have unit roots, are stationary and can be analysed in the next step.

Table 3. Basic descriptive analysis table of the log returns of companies' stocks.

Variables	Skewness	Kurtosis	JB
Stock log-Returns			
R_{BNTX}	0.583655	12.90341	p-value < 2.2e−16
R_{MRNA}	0.3952221	2.1169	p-value < 2.2e−16
R_{PFE}	−0.1294712	4.544462	p-value < 2.2e−16
R_{AZN}	−0.422983	5.080722	p-value < 2.2e−16
R_{NVAX}	1.338875	7.505994	p-value < 2.2e−16

6.3. Correlation Analysis

Financial time series tend to have relatively pronounced inertia and lags, which are manifested by the autocorrelation of the series. Therefore autocorrelation and partial autocorrelation tests need to be performed on the log returns of stocks.

Tables 5-9 show the results of the autocorrelation and partial autocorrelation tests for the series of log returns of the stocks of the five pharmaceutical companies.

Table 4. Results of unit root test.

Variables	Dickey-Fuller	p-value	Stationary
Stock log-Returns			
R_{BNTX}	−8.2847	0.01	TRUE
R_{MRNA}	−7.307	0.01	TRUE
R_{PFE}	−6.7884	0.01	TRUE
R_{AZN}	−8.2115	0.01	TRUE
R_{NVAX}	−7.3695	0.01	TRUE
Twitter Sentiment			
$B_t^{twitter}$	−4.1552	0.01	TRUE
$dB_{t,scale}^{twitter}$	−12.369	0.01	TRUE

Table 5. Results of autocorrelation and partial autocorrelation tests for BNTX.

Order	Autocorrelation	Partial Correlation
1	0.075	0.075
2	−0.113	−0.119
3	−0.123	−0.106
4	−0.074	−0.072
5	0.001	−0.015
6	−0.066	−0.098
7	−0.063	−0.074
8	0.037	0.019
9	0.116	0.08
10	−0.023	−0.059
Ljung-Box test		
X-squared	24.681	
df	12	
p-value	0.01641	

Table 6. Results of autocorrelation and partial autocorrelation tests for MRNA.

Order	Autocorrelation	Partial Correlation
1	1	0.04
2	0.04	−0.062
3	−0.06	0.056
4	0.051	0.054
5	0.062	−0.074
6	−0.075	−0.039
7	−0.049	0.02
8	0.031	−0.055
9	−0.051	−0.052
10	−0.071	0.009
Ljung-Box test		
X-squared	11.626	
df	12	
p-value	0.4762	

Table 7. Results of autocorrelation and partial autocorrelation tests for PFE.

Order	Autocorrelation	Partial Correlation
1	1	−0.156
2	−0.156	0.016
3	0.04	0.052
4	0.042	−0.033
5	−0.046	0.006
6	0.021	−0.015
7	−0.02	0.148
8	0.146	−0.109
9	−0.146	0.157
10	0.191	−0.113
Ljung-Box test		
X-squared	53.566	
df	12	
p-value	3.266e−07	

The autocorrelation of the log stock returns of the five companies is analysed according to Tables 5-9, from which it can be seen that the autocorrelation coefficient and partial autocorrelation coefficient values of the return series of

Table 8. Results of autocorrelation and partial autocorrelation tests for AZN.

Order	Autocorrelation	Partial Correlation
1	1	−0.191
2	−0.191	0.073
3	0.107	0.013
4	−0.02	−0.077
5	−0.065	0.074
6	0.095	−0.12
7	−0.156	−0.008
8	0.055	−0.106
9	−0.13	0.147
10	0.172	−0.111
Ljung-Box test		
X-squared	63.136	
df	12	
p-value	6.017e−09	

Table 9. Results of autocorrelation and partial autocorrelation tests for NVAX.

Order	Autocorrelation	Partial Correlation
1	1	0.079
2	0.079	−0.045
3	−0.039	−0.015
4	−0.021	0.122
5	0.12	0.001
6	0.023	−0.091
7	−0.096	−0.021
8	−0.042	−0.016
9	−0.001	−0.02
10	−0.011	−0.012
Ljung-Box test		
X-squared	19.019	
df	12	
p-value	0.08808	

the five companies are small, close to zero, and fluctuate up and down around the value of zero, and contain most of the AC and PAC values within the confidence interval, indicating that there is no significant autocorrelation in the log

stock returns of the five companies. Details of the ACF plots can be found in Appendix. However, when analysed based on p-values, we find some autocorrelation in the returns of certain companies. The p-values for three companies, BNTX, PFE and AZN, are less than 0.05 and it is reasonable to reject the original hypothesis and accept that there is autocorrelation in the series. The p-values for two companies, MRNA and NVAX, are greater than 0.05, which means that the series is not autocorrelated.

6.4. ARCH Effect Test

The ARCH effect is a prerequisite for GARCH family modelling and is also known as heteroskedasticity. The test is to verify whether the residuals are autocorrelated, if they are, then the ARCH effect is significant, and if not, then there is no ARCH effect. There are three commonly used tests: the squared residual correlation test, the ARCH-LM method and the graphical test.

In this paper, we use the squared residual correlation test: first build the mean model, fit μ_t, and calculate the residual $a_t = r_t - \mu_t$. The square of the residual series $\{a_t^2\}$ is used as an ARCH effect test. The Box-Ljung white noise test is performed on $\{a_t^2\}$. There is no ARCH effect when the test is not significant, and there is an ARCH effect when the test is significant. The test results are as follows.

According to Table 10, the p-values for all series are well below 0.05 and therefore highly significant, indicating that there is an ARCH effect for all five companies' log returns.

7. TGARCH-M Model Results

Since the above results show a significant ARCH effect on the series, the next decision in this paper is to use a GARCH type model to characterise the volatility of pharmaceutical company stock returns. In conjunction with the previous literature, there is an asymmetric effect of Twitter user sentiment on stock returns in financial markets, so this paper decides to choose a TGARCH (1, 1) model to conduct the study and add the standard deviation of volatility to the mean equation to reflect the relationship between volatility and returns.

The TGARCH-M (1, 1) model for each pharmaceutical company stock is set

Table 10. ARCH effect test results.

Variables	Ljung-Box test	
Stock	X-squared	p-value
BNTX	240.44	<2.2e−16
MRNA	85.286	4.01e−13
PFE	320.7	<2.2e−16
AZN	185.16	<2.2e−16
NVAX	32.453	0.001177

up as follows:

Mean equation:

$$R_t = b_0 + b_1 dB_{t,scale}^{twitter} + \tau h_t + \varepsilon_t$$
$$\varepsilon_t \mid I_{t-1} \sim N(0, \sigma_t^2) \qquad (18)$$
$$\varepsilon_t = v_t \sqrt{h_t}$$

Variance equation:

$$h_t = \sigma_t^2 = \alpha_0 + \beta h_{t-1} + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 dB_{t,scale}^{twitter} + \gamma D_{t-1} \varepsilon_{t-1}^2, \qquad (19)$$

where D_{t-1} is the indicator variable on ε_{t-1}. And when $\varepsilon_{t-1} < 0$, $D_{t-1} = 1$. When $\varepsilon_{t-1} \geq 0$, $D_{t-1} = 0$. b_1 shows the relationship between log stock returns and the Twitter user sentiment index. If $b_1 > 0$ and statistically significant, it indicates that the more optimistic the Twitter user's sentiment is, the higher the return on the asset, and conversely if the Twitter user's sentiment is more pessimistic, the lower the return on the asset. b_2 represents the relationship between risk and return. If $b_2 > 0$ and significant, it means that stock returns of pharmaceutical companies are rewarded by risk, if $b_2 < 0$ it means that stock returns of pharmaceutical companies are penalised by risk. a_2 represents the relationship between log return volatility and Twitter user sentiment. If $a_2 > 0$ and statistically significant, it indicates that Twitter user sentiment will correct return volatility in the same direction, and if $a_2 < 0$ and statistically significant, it indicates that Twitter user sentiment will correct return volatility in the opposite direction. γ indicates the coefficient of the asymmetric effect of the model. If $\gamma > 0$, it indicates that negative sentiment will have a greater impact on returns than positive sentiment at the same intensity, and if $\gamma < 0$, it indicates that negative sentiment will have a smaller impact on returns than positive sentiment at the same intensity. Tables 11-15 show the estimation results of the TGARCH-M (1, 1) model.

Table 11. TGARCH-M (1, 1) model results for BNTX.

Variables	Coefficients	Std. Error	t value	p-value
b_0	0.027225***	0.000164	165.675	0.0000
τ	−0.399048***	0.004921	−81.087	0.0000
b_1	−0.004156***	0.0001	−41.748	0.0000
α_0	0.000938***	0.000004	229.254	0.0000
α_1	0.402034***	0.001781	225.679	0.0000
β	0.291396***	0.001176	247.817	0.0000
γ	−0.286433***	0.001608	−178.107	0.0000
α_2	0.003275***	0.000012	264.267	0.0000
AIC	−2.9545***			
LogLikelihood	559.011***			

Table 11 is the TGARCH-M model result for BNTX.

From Table 11, it can be seen that. 1) From the AIC values and LogLikehood values of the model results, it can be concluded that TGARCH-M (1, 1) fits the asymmetry of the shocks relatively well. 2) The p-values of all coefficients are less than 0.05, indicating that these coefficients are significant at least at the 5% level of significance, validating the validity of the model. 3) The b_1 in the mean equation is less than 0, indicating that there is an inverse fluctuation relationship between Twitter user sentiment and BNTX stock returns, with BNTX stock returns decreasing when Twitter user sentiment tends to be optimistic and increasing when Twitter user sentiment tends to be pessimistic. 4) The return-risk coefficient τ is less than 0, which means that return is negatively correlated with its volatility, and an increase in expected risk in the market will result in a corresponding decrease in return. 5) α_2 is significantly greater than 0, which means that changes in Twitter user sentiment correct the volatility of returns in the same direction, i.e., when Twitter user sentiment tends to be optimistic, investors' blind confidence in the market will lead to an increase in risk-taking, and when Twitter user sentiment tends to be pessimistic, investors' caution in the market will lead to less risk-taking as well. 6) The coefficient of the asymmetric effect, γ, is significantly smaller than zero, indicating that positive sentiment is more likely to produce large fluctuations in returns compared to negative sentiment.

Table 12 is the TGARCH-M model result for MRNA.

From Table 12, it can be seen that. 1) From the AIC values and LogLikehood values of the model results, it can be concluded that TGARCH-M (1, 1) fits the asymmetry of the shocks relatively well. 2) The p-values for most of the coefficients are less than 0.05, indicating that these coefficients are significant at least at the 5% level of significance, validating the validity of the model. 3) The b_1 in the mean equation is greater than zero, indicating that there is a positive relationship

Table 12. TGARCH-M (1, 1) model results for MRNA.

Variables	Coefficients	Std. Error	t value	p-value
b_0	0.004037*	0.013998	0.28838	0.773057
τ	0.057001***	0.265563	79.21464	0.030046
b_1	0.002204***	0.013377	43.16477	0.019124
α_0	0.000173*	0.000198	0.87187	0.383282
α_1	0.126137***	0.046252	2.72718	0.006388
β	0.863839***	0.057181	15.10717	0
γ	−0.095087***	0.046533	−12.04342	0.021011
α_2	0.00218***	0.00089	250.32	0.000323
AIC	−2.9798			
LogLikelihood	564.1121			

between Twitter user sentiment and MRNA stock returns, with MRNA stock returns increasing when Twitter user sentiment tends to be optimistic and decreasing when Twitter user sentiment tends to be pessimistic. 4) The return-risk coefficient τ is greater than 0, which means that return is positively correlated with its volatility, and when the expected risk in the market increases it will lead to a corresponding increase in return. 5) α_2 is significantly greater than 0, which means that changes in Twitter user sentiment correct the volatility of returns in the same direction, i.e., when Twitter user sentiment tends to be optimistic, investors' blind confidence in the market will lead to an increase in risk-taking, and when Twitter user sentiment tends to be pessimistic, investors' caution in the market will lead to less risk-taking as well. 6) The coefficient of the asymmetric effect, γ, is significantly smaller than zero, indicating that positive sentiment is more likely to generate large fluctuations in returns compared to negative sentiment.

Table 13 is the TGARCH-M model result for PFE.

From Table 13, it can be seen that. 1) From the AIC and LogLikehood values of the model results, it can be concluded that TGARCH-M (1, 1) fits the asymmetry of the shocks relatively well. 2) The p-values for most of the coefficients are greater than 0.05 and the validity of the model is not very good. 3) The b_1 in the mean equation is greater than 0 indicating a positive volatility relationship between Twitter user sentiment and PFE stock returns, but the b1 coefficient is insignificant. 4) The return-risk coefficient τ is greater than 0 and insignificant. This implies that there is no GRACH-M phenomenon in returns and no risk premium. 5) α_2 is greater than 0, but is extremely insignificant. This indicates that there is no significant effect of changes in Twitter user sentiment on returns. 6) The coefficient of the asymmetric effect, γ, is less than 0, but also extremely insignificant. It indicates that the asymmetric effect is insignificant

Table 13. TGARCH-M (1, 1) model results for PFE.

Variables	Coefficients	Std. Error	t value	p-value
b_0	−0.001207*	0.003037	−0.397314	0.691136
τ	0.02423*	0.184274	0.131491	0.895387
b_1	0.005269*	0.004561	1.155153	0.248028
α_0	0.000009***	0.000001	16.833482	0
α_1	0.116183***	0.05657	2.053798	0.039995
β	0.870943***	0.042035	20.719256	0
γ	−0.028213*	0.070845	−0.39824	0.690453
α_2	9.06e−11*	0.000022	0.000004	0.999997
AIC	−5.3068			
LogLikelihood	997.7233			

Table 14. TGARCH-M (1, 1) model results for AZN.

Variables	Coefficients	Std. Error	t value	p-value
b_0	0.000524*	0.009099	0.057562	0.9541
τ	−0.040856*	0.375671	−0.108755	0.9134
b_1	0.003575*	0.012823	0.278813	0.78039
α_0	0.000076*	0.000055	1.371215	0.17031
α_1	0.324114*	0.933965	0.34703	0.72857
β	0.573558*	0.81452	0.704167	0.48133
γ	−0.169834*	0.761473	−0.223034	0.82351
α_2	8.77e−17*	0.000552	0	1
AIC	−5.128			
LogLikelihood	9,863,804			

Table 15. TGARCH-M (1, 1) model results for NVAX.

Variables	Coefficients	Std. Error	t value	p-value
b_0	0.096441***	0.000524	183.8934	0.0000
τ	−1.002087***	0.004532	−221.0986	0.0000
b_1	−0.000823***	0.000001	−1503.9104	0.0000
α_0	0.00012***	0.000001	197.8503	0.0000
α_1	0.000246***	0.000002	162.2732	0.0000
β	0.997145***	0.009568	104.2132	0.0000
γ	−0.041576***	0.000233	−178.4133	0.0000
α_2	0.002876***	0.000006	250.4324	0.0000
AIC	−1.98			
LogLikelihood	357.5261			

and that good news and bad news have the same impact on PFE stock returns.

Table 14 is the TGARCH-M model result for AZN.

The case of AZN stock is similar to that of PFE in that both companies are UK companies and both suffered large declines in March 2020, while the stocks of three other companies were rising during the same period, thus possibly causing the relationship between Twitter user sentiment and stock returns to become insignificant.

Table 15 is the TGARCH-M model result for NVAX.

The NVAX stock is similar to BNTX in terms of the magnitude of the coefficients and the positives and negatives, so the analysis of NVAX is similar to that of BNTX and will not be repeated here.

8. Robustness Test

In order to check the robustness of the results, a robustness check is required to ensure that the model is reasonable and valid, and that the results are unbiased and valid. The robustness of the model will be tested by increasing the sample size and replacing the autocovariance, based on previous robustness testing 28 Junqi Chen methods and the fact that this paper has already conducted a split-sample regression and dealt with the endogeneity problem.

Substitute Variables

Regressions were performed by replacing the variables of interest and comparing the differences in the before and after results. The SINOVAC vaccine was not included in the sample selection for this study, given that it is a large-scale and inactivated vaccine used in China, unlike the inactivated vaccines we studied. Therefore, the SINOVAC vaccine was considered to be added to the sample for regression comparison in the robustness test.

Therefore, after adding the relevant sample data and performing the above series of processing, a new investor sentiment index is formed, regressed, and the corresponding results are compared and analyzed. The specific results are as follows.

The above Table 16 shows the results of the regression with the addition of SINOVAC.

The b_1 in the mean equation is less than 0, indicating that there is an inverse fluctuation relationship between Twitter user sentiment and SINOVAC stock returns, with SINOVAC stock returns decreasing when Twitter user sentiment tends to be optimistic and increasing when Twitter user sentiment tends to be pessimistic.

The return-risk coefficient τ is less than 0, which means that return is

Table 16. TGARCH-M (1, 1) model results for SINOVAC.

Variables	Coefficients	Std. Error	t value	p-value
b_0	0.017325***	0.000123	162.275	0.0000
τ	−0.288059***	0.005032	−82.012	0.0000
b_1	−0.005267***	0.0001	−32.659	0.0000
α_0	0.000847***	0.000005	318.426	0.0000
α_1	0.301025***	0.002691	132.456	0.0000
β	0.182357***	0.001245	341.627	0.0000
γ	−0.292456***	0.001346	−125.209	0.0000
α_2	0.002345***	0.000011	235.672	0.0000
AIC	−2.8365			
LogLikelihood	542.012			

negatively correlated with its volatility, and an increase in expected risk in the market will result in a corresponding decrease in return.

Compared with the coefficients of the regression results above, there is little change and both are negatively correlated. Thus, according to the comparison of the regression results before and after increasing the sample, it can be found that the coefficients of the main variables do not change much and the direction of influence remains the same, both are positive, so the model can be considered robust and the results of the regression analysis are reliable.

9. Future Research

In future research, we should not limit ourselves to crawling Twitter data, but should consider more social platforms, such as Sina-Weibo in China and Facebook in the US. A mix of data from mainstream platforms around the world is used as a sentiment index. Although a large body of literature uses data from Twitter, other platforms can still provide huge amounts of meaningful data. In addition, with the development of web technologies, it is increasingly difficult to access data on a single platform, whether from Twitter or other platforms, so it is essential to have access to data from multiple platforms.

As of 9 July 2021, only five mainstream vaccines have been considered in this article, but as vaccine research and development is taking place in countries around the world, more and more pharmaceutical companies are undertaking vaccine development and more and more countries are starting to launch new vaccines. Examples include the Kexing and Sinopharm vaccines in China, and the Janssen vaccine in the USA. Therefore, upcoming vaccines from other countries should be taken into account in the next studies. Doing so will expand the sample and lead to more valuable and meaningful results.

The reason for not presenting the impact of macroeconomic cycles in this paper is that macroeconomic cycle data is mostly monthly data, but also quarterly data such as GDP, which is the daily data chosen for this paper. The conclusions of this paper would be more reliable if monthly data could be taken and compared to the user sentiment index which presents macroeconomic cycles.

The model chosen for this paper is based on the TGARCH model and does not explore and analyse other GARCH-type models, such as the EGARCH model. This is partly due to the abundance of GARCH-type models which cannot be compared one to the other, and partly because the model is more mature and easier to understand.

10. Conclusion

This paper proposes a validated optimisation method for constructing a web sentiment index based on previous work on the construction of web sentiment indices. The optimisation method considers optimism, pessimism, and neutral sentiment together and incorporates the number of tweets into the model. It also expresses the user's sentiment steering in terms of differentials, with the final

normalisation being that the range of variation is between −1 and 1.

This paper uses the TGARCH-M model to investigate the impact of Twitter sentiment on stock returns.

A summary of previous research findings was first presented, followed by the construction of an internet sentiment index using the new method just mentioned. The internet sentiment was mainly tweeted about Covid-19 and vaccines, i.e. Twitter sentiment. The stocks were selected from five of the best-known pharmaceutical companies, which produce five of the most popular vaccines in the world. And the time span chosen was from January 1, 2020, to July 9, 2021. A TGARCH-M (1, 1) model is then used to investigate the impact of the network sentiment index constructed in this paper on stock returns.

The results show that:

1) The TGARCH-M model fits the asymmetry of the news shock relatively well.

2) There is an inverse correlation between Twitter sentiment and the stock returns of BioTech and Novovax, in that the stock returns decrease when Twitter sentiment tends to be optimistic and increase when Twitter sentiment tends to be pessimistic.

3) There is a correlation between Twitter sentiment and Moderna's stock returns in the same direction, in that stock returns increase when Twitter sentiment tends to be optimistic and decrease when Twitter sentiment tends to be pessimistic.

4) However, the relationship between Twitter sentiment and stock returns of Pfizer and AstroZeneca is not significant. The reason for this is that both companies are UK companies and both suffered a big drop in March 2020 while the other three companies' stocks were rising in the same period, thus possibly causing the relationship between Twitter user sentiment and stock returns to become insignificant.

5) Returns are negatively correlated with their volatility, so when the expected risk in the market increases it will lead to a corresponding decrease in returns.

6) Changes in Twitter sentiment modify return volatility in the same direction, meaning that when Twitter user sentiment tends to be optimistic, investors' blind confidence in the market leads to greater risk-taking, and when Twitter user sentiment tends to be pessimistic, investors' caution in the market leads to less risk-taking.

7) Positive sentiment is more likely to generate large swings in returns than negative sentiment.

We conclude this article by urging investors to take full account of news sentiment, and the impact of internet sentiment on share prices when investing in equities. Senior investment institutions, on the other hand, should also fully explore the value of sentiment and capture the sentiment of the public in the social dimension to better their investment behaviour. The government, on the other hand, should pay more attention to public opinion, as their sentiment often

deeply reflects the sentiment of the market. The sentiment of internet users and the sentiment of the market are closely related, and by capturing the internet sentiment, they can better understand and grasp the changes in the financial market, so as to provide powerful help in formulating more accurate policies.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Ahmad, A., & Francq, C. (2016). Poisson Qmle of Count Time Series Models. *Journal of Time Series Analysis, 37,* 291-314. https://doi.org/10.1111/jtsa.12167

Ahmed, J. (2016). A Conditionally Heteroskedastic Binary Choice Model for Macro-Financial Time Series. *Journal of Statistical Computation and Simulation, 86,* 2007-2035. https://doi.org/10.1080/00949655.2015.1099159

Alexander, C., & Lazar, E. (2006). Normal Mixture GARCH (1, 1): Applications to Exchange Rate Modelling. *Journal of Applied Econometrics, 21,* 307-336. https://doi.org/10.1002/jae.849

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

Bachelier, L. (1900). Théorie de la spéculation. *Annales scientifiques de l'École Normale Supérieure, 17,* 21-86. https://doi.org/10.24033/asens.476

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

Barberis, N., Shleifer, A., & Vishny, R. W. (2005). A Model of Investor Sentiment. In R. H. Thaler (Ed.), *Advances in Behavioral Finance* (Vol. 2, pp. 423-459). Princeton University Press. https://doi.org/10.1515/9781400829125-015

Bera, A. K, Higgins, M. L., & Lee, S. (1992). Interaction between Autocorrelation and Conditional Heteroscedasticity: A Random-Coefficient Approach. *Journal of Business & Economic Statistics, 10,* 133-142. https://doi.org/10.1080/07350015.1992.10509893

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics, 31,* 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

Bollerslev, T. (1987). A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *The Review of Economics and Statistics, 69,* 542-547. https://doi.org/10.2307/1925546

Bollerslev, T., & Ghysels, E. (1996). Periodic Autoregressive Conditional Heteroscedasticity. *Journal of Business & Economic Statistics, 14,* 139-151. https://doi.org/10.1080/07350015.1996.10524640

Brown, G. W., & Cliff, M. T. (2005). Investor Sentiment and Asset Valuation. *The Journal*

of Business, 78, 405-440. https://doi.org/10.1086/427633

Caporin, M., & McAleer, M. (2006). Dynamic Asymmetric GARCH. *Journal of Financial Econometrics, 4,* 385-412. https://doi.org/10.1093/jjfinec/nbj011

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

Christian, G., & Monfort, A. (1992). Qualitative Threshold Arch Models. *Journal of Econometrics, 52,* 159-199. https://doi.org/10.1016/0304-4076(92)90069-4

Cutler, D. M., Poterba, J. M., & Summers, L. H. (1988). *What Moves Stock Prices?* (Working Paper No. 2538). National Bureau of Economic Research.
https://doi.org/10.3386/w2538

Da, Z., Engelberg, J., & Gao, P. (2015). The Sum of All Fears Investor Sentiment and Asset Prices. *The Review of Financial Studies, 28,* 1-32.
https://doi.org/10.1093/rfs/hhu072

De Bondt, W. F. M., & Thaler, R. (1985). Does the Stock Market Overreact? *The Journal of Finance, 40,* 793-805. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

Dieobold, F. X. (1986). Modeling the Persistence of Conditional Variances: A Comment. *Econometric Reviews, 5,* 51-56. https://doi.org/10.1080/07474938608800096

Ding, Z., Granger, C. W. J., & Engle, R. F. (1993). A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance, 1,* 83-106.
https://doi.org/10.1016/0927-5398(93)90006-D

Engel, R. F. (1990). Discussion: Stock Market Volatility and the Crash. *Review of Financial Studies, 3,* 103-106. https://doi.org/10.1093/rfs/3.1.103

Engle, R. F. (1982) Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica: Journal of the Econometric Society, 50,* 987-1007. https://doi.org/10.2307/1912773

Engle, R. F., Lilien, D. M., & Robins, R. P. (1987) Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model. *Econometrica: Journal of the Econometric Society, 55,* 391-407. https://doi.org/10.2307/1913242

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business, 38,* 34-105. https://doi.org/10.1086/294743

Fisher, K. L., & Statman, M. (2003). Consumer Confidence and Stock Returns. *The Journal of Portfolio Management, 30,* 115-127. https://doi.org/10.3905/jpm.2003.319925

Fornari, F., & Mele, A. (1996). Modeling the Changing Asymmetry of Conditional Variances. *Economics Letters, 50,* 197-203. https://doi.org/10.1016/0165-1765(95)00736-9

Friedman, B. M., Laibson, D. I., & Minsky, H. P. (1989). Economic Implications of Extraordinary Movements in Stock Prices. *Brookings Papers on Economic Activity, 1989,* 137-189. https://doi.org/10.2307/2534463

Donaldson, R. G., & Kamstra, M. (1997). An Artificial Neural Network-Garch Model for International Stock Return Volatility. *Journal of Empirical Finance, 4,* 17-46.
https://doi.org/10.1016/S0927-5398(96)00011-4

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993) on the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance, 48,* 1779-1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

González-Rivera, G. (1998). Smooth-Transition GARCH Models. *Studies in Nonlinear Dynamics & Econometrics, 3.* https://doi.org/10.2202/1558-3708.1041

Guégan, D., & Diebolt, J. (1994). Probabilistic Properties of the β-Arch Model. *Statistica Sinica, 4,* 71-87.

Han, H., & Park, J. Y. (2008). Time Series Properties of Arch Processes with Persistent Covariates. *Journal of Econometrics, 146,* 275-292.

Higgins, M. L., & Bera, A. K. (1992). A Class of Nonlinear Arch Models. *International Economic Review, 33,* 137-158. https://doi.org/10.2307/2526988

Jegadeesh, N., & Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance, 48,* 65-91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Kumar, A., & Lee, C. M. C. (2006) Retail Investor Sentiment and Return Comovements. *The Journal of Finance, 61,* 2451-2486. https://doi.org/10.1111/j.1540-6261.2006.01063.x

Lee, G. G. J., & Engle, R. F. (1993) *A Permanent and Transitory Component Model of Stock Return Volatility* (Discussion paper No. 9244). Department of Economics, University of California.

Li, c. W., & Li, W. K. (1996). On a Double-Threshold Autoregressive Heteroscedastic Time Series Model. *Journal of Applied Econometrics, 11,* 253-274. https://doi.org/10.1002/(SICI)1099-1255(199605)11:3%3C253::AID-JAE393%3E3.0.CO;2-8

Li, D., Ling, S., & Zhang, R. (2016). On a Threshold Double Autoregressive Model. *Journal of Business & Economic Statistics, 34,* 68-80. https://doi.org/10.1080/07350015.2014.1001028

Liu, S. (2015). Investor Sentiment and Stock Market Liquidity. *Journal of Behavioral Finance, 16,* 51-67. https://doi.org/10.1080/15427560.2015.1000334

Maheu, J. M., & McCurdy, T. H. (2004). News Arrival, Jump Dynamics, and Volatility Components for Individual Stock Returns. *The Journal of Finance, 59,* 755-793. https://doi.org/10.1111/j.1540-6261.2004.00648.x

Milhøj, A. (1987). *A Multiplicative Parameterization of ARCH Models.* Universitetets Statistiske Institut.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text Mining for Market Prediction: A Systematic Review. *Expert Systems with Applications, 41,* 7653-7670. https://doi.org/10.1016/j.eswa.2014.06.009

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica: Journal of the Econometric Society, 59,* 347-370. https://doi.org/10.2307/2938260

Nowicka-Zagrajek, J., Weron, A. et al. (2001) Dependence Structure of Stable R-Garch Processes. *Probability and Mathematical Statistics-Wroclaw University, 21,* 371-380.

Pantula, S. G. (1986). Modeling the Persistence of Conditional Variances: A Comment. *Econometric Reviews, 5,* 71-74. https://doi.org/10.1080/07474938608800099

Raissi, N., & Missaoui, S. (2015). Role of Investor Sentiment in Financial Markets: An Explanation by Behavioural Finance Approach. *International Journal of Accounting and Finance, 5,* 362-401. https://doi.org/10.1504/IJAF.2015.076182

Roberts, H. V. (1959). Stock-Market "Patterns" and Financial Analysis: Methodological Suggestions. *The Journal of Finance, 14,* 1-10. https://doi.org/10.1111/j.1540-6261.1959.tb00481.x

Robinson, P. M. (1991). Testing for Strong Serial Correlation and Dynamic Conditional Heteroskedasticity in Multiple Regression. *Journal of Econometrics, 47,* 67-84. https://doi.org/10.1016/0304-4076(91)90078-R

Sentana, E. (1995). Quadratic Arch Models. *The Review of Economic Studies, 62,* 639-661. https://doi.org/10.2307/2298081

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

Takaishi, T. (2017). Rational Garch Model: An Empirical Test for Stock Returns. *Physica A: Statistical Mechanics and Its Applications, 473,* 451-460. https://doi.org/10.1016/j.physa.2017.01.011

Taylor, S. J. (1986). *Modeling Financial Time Series.* Chichester.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance, 62,* 1139-1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

William Schwert, G. (1989) Why Does Stock Market Volatility Change Over Time? *The Journal of Finance, 44,* 1115-1153. https://doi.org/10.1111/j.1540-6261.1989.tb02647.x

Wurgler, J., & Zhuravskaya, E. (2002). Does Arbitrage Flatten Demand Curves for Stocks? *The Journal of Business, 75,* 583-608. https://doi.org/10.1086/341636

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

Zakoian, J.-M. (1994). Threshold Heteroskedastic Models. *Journal of Economic Dynamics and Control, 18,* 931-955. https://doi.org/10.1016/0165-1889(94)90039-6