ERASMUS UNIVERSITY ROTTERDAM Erasmus School of Economics

Master Thesis [Financial Economics]

‘Media Coverage Influence on Asset Prices: Empirical Assessment of Short Message Effect’

Name student: Andrey Bekarev

Student ID number: 567345

Supervisor: Prof. S. Vogel

Second assessor: Dr. G. Cocco

Date final version: 22 – 10 – 2022

Media Coverage Influence on Asset Prices: Empirical Assessment of Short Message Effect

The panel regression analysis of tweets related to the market traded companies influence over the fluctuations of these companies' stocks.

Author: Andrey Bekarev
Supervisor: prof. S. Vogel

Abstract

This paper provides the research of the short public messages effect over the companies' stocks volatility. This research is conducted using almost five years' timeframe data sample and incorporates companies with various properties though similar in the way that most of them occur in a social media coverage that drives their stock volatilities.

The research is conveyed incorporating the data from 45 different companies within the period starting in the beginning of 2018 and ending in the second quarter of 2022. The total number of observations accounted for 41741. Within the given sample I found out empirical proof that tweets influence the performance of stocks of targeted companies. The effect is observed in terms of volatility of the stock, but not its returns. In order to find this correlation, there was a new generic variable created that absorbs the effects of time and tweet count fluctuation. Still the results are controversial in both within the company types and the within time periods. Thus, it brings us to the new frontier of the research in the time of the enormous information disturbance that influence our day-to-day activities.

This research basics might be used in the quantitative arbitrage trading models that assess and predict the fluctuation from median.

Thesis version: October 2022 (Final Draft)
Key words: Stock volatility, social media coverage, media influence, tweet count, trader attention span, retail investor

Acknowledgements

I would like to address my gratitude to my Thesis supervisor prof. Sebastian Vogel for his thorough review and critical assessment of the work that helped me to find the most vivid intricacies of the modern social media and the stock market. The new era of the trading and asset management that we are entering barely would avoid the magnitude of the public opinion expressed as short sparks at a glance barely noticed. As well as I would like to thank my second assessor dr. Giovanni Cocco for taking the time to assess the paper.

NON-PLAGIARISM STATEMENT

By submitting thus thesis, the author declares to have written this thesis completely by himself/herself, and not to have used sources or resources other than the ones mentioned. All sources used, quotes and citations that were literally taken from publications, or that were in close accordance with the meaning of those publications, are indicated as such.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Table of Contents

Chapter 1: Introduction

1.1 General issues

Contemporarily activities could rarely be observed without a social media. These days we are not just pure consumers but also creators of the news ourselves. Furthermore, social media coverage is crucial not just for our daily activities, but also in terms of asset pricing. It might be considered that regulation of the stock market could magnify or mitigate such an effect of the media coverage on the stock market.

One of the most prominent episodes on the stock market of the previous year was a campaign to support GameStop share prices organized among retail investors through social media (Reddit). This has led to the GME price skyrocketing fivefold and the following billion losses of several hedge-funds betting short. This is one of the 2021 brightest examples of how social media context might impact the market.

Of course, there are a lot of factors that determine a return on an asset and its risk which is associated with volatility. Moreover, there are a list of parameters which reflect the private investors' behaviour.

The private investors' behaviour is a group's phenomenon. So, we can assume that their behaviour is being influenced by the news, social media messages and rumors, to generalize, the media coverage. Nowadays investors discuss alluring investment scopes in social networks such as Twitter and others. The fact of being noticeable reflects an investment appeal in some degree.

What is important for a company to became noticeable in social networks? Despite the common knowledge it is not enough in an economic sense to become noticeable just by pumping the news space with the company news. Within the research it was identified that not only frequency of the news coverage play role, but also the connotation of the news. Thus, the sentiment should be feasible in terms of the asset price what was noted by Noah Mukhtar in his short article[1]. Many other researchers also claim that sentiment is a significant measure, unfortunately it was not proved with the data collected. Quality of news – native precipitation and thus the quality of the source also gives the ground and magnitude while setting up the investors financial decision making as far as the timing. In the world that generates information flow on a progression scale it is now way more difficult to follow the news than it was ten years ago. And, perhaps, it might be even more difficult ten years from now as the density of the

[1] https://towardsdatascience.com/can-we-beat-the-stock-market-using-twitter-ef8465fd12e2

information flow would multiple several folds. Thus, to catch a trader with the right news hint while he or she is on the way to the desk might be a crucial point.

Moreover, some other factors can play a crucial role for the private investors' decisions. These factors could be more classical as they incorporate quantitative economic measures as the price of the stocks to fit into investors portfolio, business goodwill, size of the company and the value of a brand. Nevertheless, important parameters might be an ESG keen of the company and leader's values.

The main goal of the Master thesis is to find if the media coverage influences the companies' share price or any indices of their share prices, returns or volatility.

The goal of the research determines the objectives. Within the main objectives there were a thorough research of the papers that were covering the social media and the trading topics to see what were the core investigations methods and objectives to incorporate those results in the Thesis. Furthermore, after collecting the data, there was an aim to construct the econometric model that defines the economic sense with the statistical ground for this research. Several side objectives included the test of the practical matters such as an ETF made on meme-stocks.

The objects of the research are two types of companies with the following features. From the author's point of view, the big companies, which are famous and stable, and the new companies, which are hyped and prominent, are mostly influenced by the media coverage factor.

The subject of the research is the way of the media coverage influence on asset prices that can be determined by the private investors' behaviour.

Chapter 1 is the Introduction to the research that covers the ideas that explain the relevance of the topic. This chapter is devoted to the review and analysis of the scientific articles in order to find the gaps to investigate. Chapter 2 is devoted to the theoretical background and extended evolution of the efficient market hypothesis and behavioral finance. Chapter 3 describes the methodology of the research and statistics used. Chapter 4 discusses the results of econometric research evaluation. Chapter 5 is dedicated to further research and depicts the conclusion of the current paper.

1.2 Literature review

The research works published are being fulfilled with analyses of how the public media coverage influences asset prices. Furthermore, there are scientific papers that analyze the social media influence on the different types of assets. The most relevant research papers were listed in Table of Appendix 1. The articles are compared in accordance to such characteristics as method used, objects under investigations, data and results.

The thesis was primarily incorporated by the papers that guided us with the course of research, variables that should be testified and the methodology. In the most of works panel regression analysis was especially used that was the basic reason to choose a cluster of companies to analyze. Then the central course of research showed that media generally influence the stock market. From retrospective point of view research papers depicts that earlier exactly large articles in a specified source were the objects of research, but recently small messages and the social media appeared under the scrutiny research of economists. The next clue from the papers was the variables that we consolidated in the thesis. As most of the works tend to take volatility as the dependent variable, we took it as well. Still it is not crystal clear if there is a robust connectedness to the returns or mainly to the volatility and thus we decided to test returns simultaneously. Researchers also gravitate to take not just pure tweet counts into the model, but the sentiment colour of it. Unfortunately, in the thesis we were not able to dig this deep and analyze our data on our own in terms of connotation, thus we used the sentiments given by Bloomberg database and generated our own variable that might appeared to be a quite robust proxy – LDT (a lagged deviation of tweet count).

After the analysis of the list of articles in Table of Appendix 1, it might be inferred that research of social media coverage in concern with asset prices has intensified lately. In the field of research there is a lot of papers published regarding the social media influence on stock market, crypto assets and indices. Authors of the articles listed in Table of Appendix 1 notice the social media coverage impact on prices of the assets (Chen et al., 2013; Yang et al., 2019), firm value (Wong, 2021), returns (Dyck, 2003; Freiberg et al., 2018; Huan, 2018), volatility (Haroon, 2020; Jiao, 2020; Umar et al., 2021) and turnover (Jiao, 2020, Ben-Rephael, 2022).

These authors explore single stocks behavior as far as the joint cluster of companies in terms of returns, volatility and turnover. In the scope of interest were such companies which grow and develop faster than competitors in the field with an investment opportunity (Bushma, 2019). These companies were under analysis within distinctive economical periods and conditions (Fraiberg, 2018). Furthermore, companies that undergo internal or external alternations like M&A were evaluated (Ahern, 2015). The data used for the analysis is highly divergent. Previously the economic articles were used by the authors as the statistical data from well-known publishers or journals (Dyck, 2003; Chen, 2013; Fraiberg, 2018). Though during

2020-2021 researchers started using specific indices estimated by media platforms based on the articles and publications stored within.

The method of research is a ground to any research paper. In most articles' authors used econometric approach as a panel data analysis (Chen, 2013; Ahern, 2015; Yang, 2019), cross section analysis (Huan, 2018; Bushma, 2019) and GARCH modelling (Haroon, 2020). In addition, it is worth to mention the paper (Su et al., 2021) that discloses the methodology used by authors to assess the impact of the media on the price of a stock.

For the purpose of the current paper the most recent researches were used. Mainly because of the up-to-date context that involves uncertainty and higher risks on the market and furthermore the authors of the articles that imbed volatility as one of the main risks estimate within the models. This parameter reflects investors' perception of the market as well as the magnitude of uncertainty of the market. Furthermore, the excess volatility on the market within recent years might be associated with the inflow of retail investors on the market and thus chaotic actions of them.

For example, the research (Umar et al., 2021) mostly covers the analysis of investors behaviour on the stock market during Covid-19 pandemic, when mass short selling occurred. Authors "examine the returns and volatility connectedness between media coverage index (MCI) and high short interest stocks during the recent Covid-19 pandemic". Returns are testified to be affected by the media coverage. Authors of the other work (Wong and Zhang, 2021) "examine the value relevance of corporate reputation risks (CRR) from adverse media coverage of environmental, social and governance (ESG) issues on stock performance". The paper rationalizes negative skewness of investors preferences.

Modern papers' authors of 2020-2021 tend to investigate the volatility of stock prices (Haroon, 2020; Jiao, 2020; Umar, 2021; Ben-Rephael, 2022). This dense research scrutiny also pushed the course of research of this paper. These papers covered several parameters of the media as the rise of mentioning and sentiment magnitude that drives volatility further and keeps the notice of high magnitude of media coverage on the risk and return on the stock market.

Chapter 2: Theory and Hypotheses

The core research of the factors influencing the asset prices provide the theoretical solid ground for this Master thesis research. One of the most issues sound persists the question of what determines the prices of the assets. Robert Shiller and Eugene Fama research especially this field and received Nobel Prizes in 2013 (Gelman and Sprenger, 2014). Other Nobel laureates such as Sanford Grossman, Joseph Stiglitz, Daniel Kahneman and Richard Thaler also work in the area of behavioral and socio-economics.

This chapter of the Master thesis covers the brief overview of the evolution of asset pricing, incorporating the information factors. E. Fama tends to be the proponent of the efficient market hypothesis that claims the prices of the assets to incorporate swiftly all relevant information. Grossman and Stiglitz have theoretically proved that the prices of the assets fluctuate under the incoming new information. Random walk methodology helped them to adjust the theory with the assumption of random new information inflow. Robert Shiller claimed that “investment coefficient” relevant to the asset could be alternated sharply under the investors’ perception of the circumstances. Furthermore, Daniel Kahneman is considered a founding father of the behavioral economics that is being currently used to analyze investors decisions within the financial markets; and Richard Thaler follows some of his ideas in behavioral finance.

2.1. Efficient market hypothesis

The hypothesis is a core to the modern finance, illustrating the asset pricing effects on the market. Efficient Market Hypothesis (EMH) asserts that market prices fully reflect the open information about the assets to the investors that might affect the asset price.

Fama is one of the authors of the hypothesis (Fama, 1970). He believed that if there is no possibility to predict the price of the asset in future, there is a need to determine the variable that would be the best proxy for it. Such a variable was generalized as “information”. This factor was considered as an aggregate of the data that might affect the price of a stock. Through information, investor allegedly can validate the future price, compare it to the current price of the asset and make a decision on what to do with the current stock. Thus, the quality and the integrity of the information determines the precision of the price forecast.

There are three main types of information, based on integrity:

1) weak form – prices illustrate any information about previous events. There is no opportunity to exploit anomalies and receive abnormal returns and the price movements are determined by random walk;

2) semi-strict form – prices depict not only the previous events, but also current developments;
3) strict form – prices incorporate whole set of information, starting from the previous events, current developments, company data and insiders' information.

EMH is regularly criticized due to collisions with the real-world developments on the stock market. Though it is mostly the theoretical instrument for the analysis of the asset pricing, but still one of the core active asset management theories.

Fama provides ground in his research that the market is highly efficient while weak and/or semi-strict information is in reach. Thus, it is not possible to have robust returns on the market, exploiting only publicly available information. Fama (1991) admits that prices do not incorporate all possible information as it would be a controversy to the research of Stiglitz and Grossman (1980). Prices illustrates the information at the such a rate that costs of mining the information equals the conceivable profits that could be gained through that information.

Sanford J. Grossman and Joseph E. Stiglitz (Grossman and Stiglitz, 1980) showed that it is impossible for a market to be perfectly informationally efficient. Because information is costly, prices cannot perfectly reflect the information which is available, since if it did, investors who spent resources on obtaining and analyzing it would receive no compensation. Thus, a sensible model of market equilibrium must leave some incentive for information-gathering (security analysis) (Sewell, 2011).

2.2. Asset price volatility

Robert Shiller has also based his researches on the EMH assumptions, though the object of research was different. The limits of the price fluctuations were his objective instead of the price forecast (Shiller, 1981). Before Shiller published his research, it assumed that dividend pay off information was the key factor in terms of price fluctuations of the stocks.

If the market is efficient with no arbitrage opportunities, the price of the asset should be equal to the discounted cash flow of the asset. In case of the stock, the DCF is determined by dividend flow. Thus, dividends should be key predictors of the price of the stock. Shiller made the research within the long range of observations for stocks within the S&P500 and Dow Jones index.

The discovery by Shiller illustrates that the high variance of fluctuation of the stock price is not explained by the news about dividends. The high volatility of the stock prices is in line with the EMH only if the discount factors alternate drastically in time. In the meantime, the model is not possible to be tested straight with alternating coefficient in time as the estimates are not observed. Though Shiller was able to defend the case that the volatility of the discount factors should be inexplicably high in order to be in line with the model. Hence, Shiller was highly skeptical about the fact that in a model with rational investors it is plausible to have such a high variance for discount factors. On contrary, he doubted the statement that investor is rational in economic model (Shiller, 1981)

2.3. Behavioral finance

Behavioral finance theory is the logical successor for the financial and asset pricing theories. The most prominent researchers in the field are Daniel Kahneman and Richard Thaler. Kahneman observed that economic actors making decisions in most cases are irrational in economic terms. The decisions of these actors are made swiftly, considering only shallow factors and incorporating historical data. Such a decision making processed is developed in our nature through fundamental psychophysiological mechanics in our brain. This presumption has paved the ground for the research of Kahneman, who developed and verified through field tests his perspectives theory (Kahneman and Tversky, 1979). The key point of the theory is the risk averse behavior.

This theory claims that losses of the amount of assets is considered more sufficient, than the gain of the same amount of money. In the other words, investors suffer losses at a higher magnitude than theoretical perspective of the gain of wealth.

Most of the people are risk averse even if the possibility of the loss is essentially low. The reason behind this pattern is that how people estimate uncertain possibilities and thus how they calculate the outcome on the given possibilities.

Richard Thaler has developed “cognitive account” theory. This theory explains how individuals design separate accounts in their mind and focus mainly on individual cases than on the summary of them. Furthermore, Thaler depicted the model of how people tend to follow sparkling temptations what explains the inability of the majority to make savings (Thaler and Benartzi, 2004).

In addition, Thaler claims that a lot of investors decisions are guided by the emotions and the current mood. In field research clarify that it might be way more frequent cases than majority assumes. To illustrate, investors risk averse might be altered by the way how the strategy is drawn, the state of hunger in the moment, the weather outside and furthermore. Lots of these factors and beyond affect the decision-making process even if the investors don’t see it.

Individual investors behavior on the market might have the most feasible impact on the volatility on the market within the current research timeframe.

Considering the amount of information circulation and the amount of it adding on the web, it is quite obvious that a human being is unable to incorporate it all and incorporate it into the decision-making process. Thus, each separate trader or economic actor would do it with its own level of scrutiny. Furthermore, taking into account the background and the capacity of each, it is wise to consider that the decision might not always be made in line with the economic perception and thus might vary. On the same stack of information, one could buy an asset, while the other will sell it. Respectively, the information flow should alternate prices, though the direction might be unclear. In other words, taking the information we have, if there is any correlation, the volatility should be influenced, but the price itself and thus gains might be blurry.

This gross assumption is guided clearly within the behavioral finance that states the economic actor as an irrational being guided by emotions. It is not hard to anticipate that if the trader receives controversial information within several minutes and he or she needs to draw a decision based on the incoming flow, the risk of failure and wrong decision-making process is higher than if the information flow is concise. According to the data gathered, we observe that when there are spikes in the mentions of any company it rarely goes one direction. In most of the cases negative and positive connotation of the messages grow simultaneously, that leads to uncertainty.

2.4. Hypotheses of the research

Hypothesis 1

Tweets influence such derivative indices on share prices as return and volatility.

The intuition behind this hypothesis is that the more company is being mentioned in the media, the more it's assets should be traded. Furthermore, it is not clear with what magnitude it would affect asset price, but every time the name or alias of the company is being visible by prospective trader, it calls for some action. And these actions might be to search for more information or directly – to sell and buy the stock.

Hypothesis 2

Media coverage influences share prices according to the behavioral finance principles that are reflected by the different investors' reaction on positive and negative tweets.

According to behavioral research, most of the investors suffer losses with the higher magnitude than feeling gains. Such risk averse hypothesis of the investors' behavior brings this paper to infer similar assumptions on the sentiment of the news. In addition, several authors tend to claim that exactly the magnitude of the sentiment of the news plays the key role on the effect inflicted.

Hypothesis 3

Tweets influence share prices with a lag of a very short period, namely 1 day.

This hypothesis is mostly driven by the attention span of the up-to-date user. It is now assumed that median attention span of the internet user is around 8 seconds. Thus, the presumption to this hypothesis is the modern life tendencies, that also include overload of the information and thus peculiarities of the human brain functioning.

Hypothesis 4

The power of the media coverage influence depends on the features of companies.

The features of the companies are illustrated in Appendix 2, though this list is not exhaustive. Key assumption is that the same number of tweets might have different impact on the stocks that have different classical parameters as volume, price, liquidity and others.

For instance, Tesla might have been under way higher social media influence due to its CEO actions than Daimler or BMW and thus the Twitter impact on Tesla has more explanatory power in the price fluctuations.

Hypothesis 5

The share price is crucial for a retail investor who has a restricted budget for their investments and is an active user of social media.

Portfolio theory claims that any investor descends his or her risk through diversification. If this is true and the rational investor exploits the same principle, thus he or she should be restricted by the budget and diversification limit. It is asserted that at least ten stocks should comprise the portfolio. Considering the median investment budget of 1000 USD, it is plausible to draw single stock price limits to test.

Hypothesis 6

The share price volatility depends on the frequency/number of tweets in social media.

This volatility hypothesis based on the clue that if the company is mentioned every day at almost the same rate and this rate might be either high or low, but it is in the line with what company usually receive – it makes no difference to the information fluctuations around it.

Hypothesis 7

The share price of the young companies from the list of ETF based on "the media coverage dependance" is influenced by tweets more than "blue chips".

This hypothesis has the intuition behind that blue chip are mostly traded by institutional investors at least in term of turnover. Thus, hype news should not influence these stocks with the same magnitude as the companies consisted of retail investors money.

Chapter 3: Research Method

3.1. Methodological approach

In order to determine if the media coverage factor influences the asset prices, as well as to assess the degree and some other details of the impact of the factor, the study was organized the way depicted in Figure 1.

Figure 1: Steps of the thesis research: methodological scheme

1. Making analysis of the scientific papers on the topic to select/determine factors affecting the asset prices, return and volatility
2. Formulating hypothesis of the research
3. Making a case for selection companies for analysis
4. Structuring of econometric model and creating a special variable that explains the media coverage influence
5. Data collecting for the companies under investigation, media coverage and financial market indices
6. Econometric estimation and making analysis of the results

Based on the analysis of the scientific literature in Chapter 1 and the theoretical background in Chapter 2 the hypotheses of the Master thesis were formulated. The analysis allows to conclude that as for the factor influencing the asset prices the information set has been increasing its value for more than 50 years. Nowadays the media coverage plays crucial role in pricing, especially in short term period and for some types of companies and their assets.

The companies for the research were selected based on their probable higher appeal to the investors and, because of this fact, a higher share price response to the news or tweets. From our point of view, they are the new and young firms, or startups, and the large firms, or well-known corporations.

In order to find the significant factors for the asset prices and their degree of influence, the econometric method panel data analysis was used. The general case of econometric model is as follows:

$$\text{Dependent variable}_{\mathrm{j,t}} = \beta_0 + \sum \beta_i factor_{i,j,t} + \varepsilon_{j,t}$$

where:

- a dependent variable is supposed to be one of the indices reflects the share price movements and results, namely, return and volatility;
- a factor is one of the variables from the list; in the case of each econometric dependence there are the variables determined by the hypothesis chosen;
- β is a coefficient;
- ε are residuals.

One of the econometric equations has the following way:

$$\text{Volatility}_{\text{j,t}} = \beta_0 + \beta_1 \text{TWC}_{\text{j,t}} + \beta_2 \text{NTWC}_{\text{j,t}} + \beta_3 \text{PTWC}_{\text{j,t}} + \beta_4 \text{LIBOR}_{\text{j,t}} + \beta_5 \text{LagTWC10}_{\text{j,t}} + \\ + \varepsilon_{j,t}$$

It reflects the dependance of the share price volatility from the variables such as number of the total, positive and negative tweets in a social network, one of the key short-term rates on the financial market, and the variable of one day lag of a deviation from 10-day moving average of general tweets.

After the thorough research of divergent variables within the data panel, the model was adjusted to the following form:

$$\text{Volatility}_{\text{j,t}} = \beta_0 + \beta_1 \text{LDT}_{\text{j,t}} + \beta_2 \text{LIBOR}_{\text{j,t}} + \varepsilon_{j,t}$$

where:

- Volatility is a volatility of the asset calculated as highest price minus the lowest price during the trading session divided by the open price;
- ß are coefficients within the regression;
- LDT is one day lag of a deviation from 10-day moving average of general tweets;
- LIBOR is the overnight rate of LIBOR;
- ε are residuals.

The ideas about the data and statistic collection details will be described in the 3.2.

Chapter 4 is devoted to the results. Every hypothesis is implied to be investigated with the help of an econometric estimation. So, having the results we can accept or reject each of them.

3.2. Data

The data sources are the Bloomberg database and the Roundhill Investments fund for the companies from the “meme index ETF”. The number of companies under investigation is 45. The list of the companies is given further in the text. The brief characteristics of them are in Appendix 2.

Initial data for media coverage are the tweet count, negative and positive tweet count. Other data are concerned with the companies and financial market conditions. They are as follows: the share prices, the volume traded of the shares, and short-term interest rates. All of these data were reviewed for the period from 9 February 2021 to 3 May 2022, for some companies the statistics was available from 10 January 2018. Data was collected daily. In total, the number of observations was 41,741.

In order to check hypothesis that tweets affect the stock market, the companies from so-called meme index ETF were chosen. That is an ETF that consists of 25[2] US companies chosen by the principle that these companies are more sensitive to the social media fluctuations than the others. This ETF was set up on 9/02/2021 by Roundhill Investments[3]. Undoubtedly, the composition of index might change within the time period, but my selection was based on the original list of 25 companies. Only two companies were omitted as there is almost no data associated with them – VIAC (ViacomCBS) and RIVN (Rivian Automotive). As of today, this index has been resorted.

The list chosen is the following: AFRM, AMC, SPCE, GME, DKNG, RBLX, SQ, HOOD, ROKU, WISH, PLTR, SNDL, TLRY, PTON, BYND, SAVA, SOFI, CLF, BBBY, BB, CCL, LCID, DWAC.

In addition to these meme index companies, there were the additional 22 companies that, in my opinion occupy social media space on a daily basis and thus should be also influenced by social media. The list of non-meme companies is the following: AAPL, AMD, AMZN, ATVI, BABA, BBY, COST, EA, FB, JD, MCD, MRNA, MSFT, NFLX, NKE, NVDA, PFE, T, TSLA, TTWO, WMT, ZM.

The following data was collected for each company from the Bloomberg database: open price, close price, high and low price; volume traded, general tweet count, negative tweet count and positive tweet count. All the data was collected on daily basis starting 10/01/2018 and ending 03/05/2022.

[2] https://www.investors.com/etfs-and-funds/sectors/sp500-meme-stock-crash-costs-speculators-48-9-billion/

[3] https://www.roundhillinvestments.com/etf/meme/

Chapter 4: Results and Discussion

4.1. Preliminary data processing

In order to process the raw data and transfer it to the data with economic sense there were additional variables generated:

1) return by subtracting close price of the current day from the previous,

2) return in % by dividing the return by the close price of previous day,

3) volatility by subtracting low price from high price and dividing by the open price of the same day,

4) for each tweet type there were generated moving averages (MA) of 10 and 100 days

5) for each tweet type there were generated a deviation from MA's by dividing absolute count to the MA for the same day. *Intuition behind this number is to see the fluctuation from the general activity in social media regarding each company.*

6) for each deviation variable and absolute variable there were additional lag variables added for 1 and 3 days.

All these data were extended by LIBOR and SOFR daily rates and in addition LIBOR for 1, 3 and 12 months. The intuition behind these coefficients is that the market takes into account the price of the investments whether on daily basis or within a longer time period.

Furthermore, these data were supplied by additional dummies: “hype” and “user”. The first dummy is ‘meme’ that defines if the company has been in the meme stock index or not. Hype was used for companies like Zoom, Pfizer, Moderna and Tesla – so the companies that were pumped regularly on social media or that were affected by some exogenous pressure. In addition, within the list of “hype” were included companies of a high upside future expectations, so basically companies like Tesla, Lucid Automotive or Digital World Acquisition Company. These companies are characterized mostly with the idea, not the real performance, so intuition behind such division is that users have extraordinary expectations referring these companies. User list was created for the companies whose business is based mostly on digital infrastructure and thus users that stroll around their infrastructure – Amazon, Facebook (Meta), EA, Activision, Microsoft etc.

With all these data a big panel dataset was constructed, incorporating all 45 companies with all the variables listed above. The list of the variables is presented in Appendix 3.

Most of the variables were tested and unfortunately there is no significant correlation with returns and any kind of tweet's variables, but volatility as a dependent variable shows sustainable, though quite low magnitude of causality with tweets.

Further subchapters are devoted to the hypotheses which were tested.

4.2. Dependent variables

There was a preliminary research in order to pick up the most significant dependent variable from the dataset that was collected. Thus, returns and volatility were under scrutiny investigation. Unfortunately, the returns in any variation showed no significant correlation with the tweet social media data. Returns have no robust dependence on the tweets, sentiments or derivative variables from the social media coverage. In the meantime, volatility shows sustainable correlation almost in every set of data collected. This has led us to omit the returns variable and keep the research within the framework of how social media influences volatility of the asset.

Short clue, why the returns seems irrelevant in our research might be that the social media noise only create uncertainty, but doesn't move the market in some specific direction.

4.3. Positive and negative tweets influence

Omitting inefficient trials, there is a panel regression outcome with volatility as a dependent variable; tweet count, **positive and negative tweet** count and one day lag of a deviation from 10-day MA of general tweets; and daily LIBOR. With all these data random-effects GLS regression estimation was made; an outcome is presented in Table 1. The table summarizes the econometric estimation results for the model with "volatility" as a dependent variable. Factors in the first columns are independent variables such as TWC – tweet count, NTWC – negative tweet count, PTWC – positive tweet count, LDT – 1-day lag deviation of 10-day MA of general tweet count, LIBOR_ON –LIBOR overnight. The only two of the factors are significant, LDT and LIBOR_ON.

Robust standard errors observed in the model represent the unbiased standard errors of the OLS coefficients under heteroscedasticity. The model adjusted to the empirical based variance of standard errors allows us to assess the model more thoroughly and precise that is necessary as our R-squared parameters are quite low.

The R-squared parameters show how much of the dependent variable variance is explained by the independent variables. R-sq within characterize the explanatory power of the median for each cluster (each of the 45 companies in our case). The parameter equal to 0.1190 shows that significant factors of the model still explain 11% of the variance of the dependent variable. In terms of how volatility is explained by the derivative variable of the number of tweets, the parameter is economically reasonable and worth the following research. Unequivocally, volatility is influenced by the number of other factors and still the proxy for tweets chosen in the research has high explanatory power. R-sq between shows how big is the difference between observable clusters, while R-sq overall shows how much of the variance of the dependent variable is explained through the variables of the model included.

Wald chi2 criterion illustrates the overall quality of the model. The Wald chi2 parameter of 46.16 claims the model significance. It depicts that the coefficients of the models all together are significant.

Sigma u is an individual cluster mean variance that depicts the individual effects influence. The lower it is, the better is the inside cluster model. In the current model it shows that the variance of the error is around 2% around the mean. Sigma e shows the mean variance of the standard errors of the model as a whole, thus the lower this number, the less variance has the model errors. The overall errors have 3.7% variance around the mean, that is higher than that inside the cluster and what once again proves that the data within the cluster has its robust individual means.

Table 1: Positive and negative tweets influence on volatility as a dependent variable

Factors	Coefficients	Robust standard errors	P-value
TWC	0.0000	0.0000	0.157
NTWC	-0.0000	0.0000	0.251
PTWC	0.0000*	0.0000	0.052
LDT	0.0003***	0.0001	0.003
LIBOR_ON	-0.0055***	0.0014	0.000
cons	0.0502	0.0040	0.000
Wald chi2		46.16	
Sigma_u		0.0203	
Sigma_e		0.0373	
R-squared within		0.1190	
R-squared between		0.2958	
R-squared overall		0.1220	
Number of observations		41,745	
Number of groups		45	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Here we may observe that tweet count, negative tweet count and positive tweet count are **insignificant within this panel**. Though 1-day lag deviation of 10-day MA of general tweet count is significant, although has quite low beta. Exactly this independent variable will be taken as the key research variable (mostly because the rest are insignificant).

About the key independent variable:

1-day lag deviation 10-day MA general tweet variable will be marked as "LDT" further in text. Intuition behind this variable I have briefly covered above in the text, but to disclose it further

– the main point is that we need to see spikes in social activity and thus especially deviation should be considered as a key independent variable.

In the same time, 100-day MA deviation is insignificant, perhaps due to the fact that modern trader/user is unable to feel this deviation within such a long period, while 10-day MA is more feasible.

Moreover, negative and positive tweet synthetic variables do not show any significance, what is slightly counterintuitive. We assume that exactly connotation or sentiment should play a key role as the market should watch precisely if news is bad or good, but according to data collected, there is no significant effect. This might be associated with the data quality as there is no insight how Bloomberg collects data and how it defines tweets were negative or positive. Furthermore, I've found the scientific paper by the research group from McGill university under the Noah Mukhtar supervision has detected[4] strong causalities of the tweet sentiments on the market. They conducted their research just with 4 companies, but they used their machine learning algorithm to define the level of the tweet sentiment. Perhaps, the key is in the quality of their ML algorithm that is more sensitive than the same in Bloomberg.

To conclude, the 1-day lag seems to be more robust than 3-day lag, while 2-day lag shows almost no significance. This also is quite intuitive as the market should fully incorporate the social media fluctuations only after it is done – so in the end of the day. The trading daily session ends while tweets still could be posted, thus it is quite consistent to have a 1-day lag.

4.4. Lagged tweet factor (LDT) influence

As in Table 1 it was illustrated that the only LDT is significant, thus all insignificant variables were omitted in order to construct a regression. The result of estimation is shown in Table 2. Within this table it might be observed that overall significance of the regression has dropped (Wald chi2 decreased to 17.33, R-sq parameters also were lowered), but the regression betas are significant and the robust standard errors are on the same level. LDT coefficient has adjusted to the 0.0004. This adjustment of the parameter gives us a clue that if even a volatility in a long term has a robust correlation with the proxy of tweets, there is a solid ground for the research to build a sentiment model for traded stocks, considering several restrictions – whether these are the types of the companies or the timeframes or additional parameters that we still have to find out.

[4] https://towardsdatascience.com/can-we-beat-the-stock-market-using-twitter-ef8465fd12e2

Table 2: LDT influence on volatility as a dependent variable. Full set

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0004***	0.0001	0.009
LIBOR_ON	-0.0065***	0.0018	0.000
cons	0.0550	0.0045	0.000
Wald chi2		17.33	
Sigma_u		0.0208	
Sigma_e		0.0392	
R-squared within		0.0273	
R-squared between		0.3855	
R-squared overall		0.0425	
Number of observations		41,745	
Number of groups		45	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

With this regression we might observe the basic correlation of the variables and thus there should be conducted further tests to determine additional parameters of the model.

According to the estimation of the regression we can conclude that the 1-day lagged tweet count variable (LDT) is significant and positive, and LIBOR interest rate is significant and negative for volatility as the dependent variable. The coefficient of the LDT variable means that in case of lagged tweets index is rising on one unit the volatility is rising on 0,04 percent. In case of LIBOR is rising on 1 per cent the volatility is falling on 0,65 percent.

Wald test proves the significance and adequacy of the model. R-squared suits for such types of the regressions.

The regression was estimated for the total sample, i.e., 45 companies and 41,745 observations for the total period under research.

4.5. Testing dummies

The next step is to test **dummies** created for the purpose of my investigation: **meme**, **hype** and **user**. The result of estimation is shown in Table 3. We have decided to search for additional determining variables of the estimated models and also to search for patterns that are being determined by the sample year. The outcome shows that **hype** and **user** dummies are **insignificant** at 5% significance level. **Meme dummy is significant**; thus, such a division is coherent. Furthermore, year dummies demonstrate that almost each year data sample has significant differences from one another. Thus, in 2020 and 2022 tweet proxy has approximately 1.5% more influence on volatility in general than in 2018 and 2021. These digits then pave the ground for the further thorough research based on the timeframe.

Table 3: Dummies influence on volatility as a dependent variable

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0004***	0.0001	0.007
LIBOR_ON	-0.0004	0.0012	0.715
HYPE	0.0334*	0.0182	0.068
USER	0.0011	0.0046	0.818
MEME	0.0064***	0.0064	0.000
Year 2019	-0.0051***	0.0018	0.004
Year 2020	0.0132***	0.0032	0.000
Year 2021	0.0054	0.0036	0.138
Year 2022	0.0158***	0.0035	0.000
cons	0.1710	0.0035	0.000
Wald chi2		277.16	
Sigma_u		0.0135	
Sigma_e		0.0387	
R-squared within		0.0504	
R-squared between		0.6875	
R-squared overall		0.2178	
Number of observations		41,745	
Number of groups		45	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

4.6. Prices of the stocks influence

In addition, within Stata program there were generated dummies based on the prices of the stocks: 1) for stocks under **25$**, 2) between **25$ and 75$**, 3) over **75$**. The idea behind this division is that **retail investors** who focus on tweet information should be strictly limited with the budget – not more than 1000$. Assuming the diversification concept we might presume that each asset should not exceed 70-100$ for such user. Furthermore, if our investor is rational, he or she should not keep in the portfolio such an asset that is indivisible, namely at least 2 stocks of each. So, we need to check this rational assumption that might not be true within the real retail investor on the market. The result of estimation is presented in Table 4:

Table 4: Share price influence on volatility as a dependent variable

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0004***	0.0001	0.006
LIBOR_ON	-0.0070***	0.0017	0.000
P_l	-0.0048	0.0076	0.531
P_m	0.0075**	0.0030	0.011
P_h		0 (omitted)	
cons	0.0542	0.0053	0.000
Wald chi2		24.49	
Sigma_u		0.0176	
Sigma_e		0.0390	
R-squared within		0.0354	
R-squared between		0.0742	
R-squared overall		0.0223	
Number of observations		41,745	

Number of groups	45

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

The most **significant** coefficients are shown for assets **between 25$ and 75$**, stocks under 25$ seems to be irrelevant, though it might be not much of the observations within the sample and thus it is better to keep this dummy for the further research. The rest seems to be insufficient.

4.7. Number of tweets influence

The next group of dummies that were created in Stata program is the tweeter count dummies. The objective to create these groups is to see if there is threshold of the number of tweets that start to affect the volatility of the asset. The result of estimation is presented in Table 5:

Table 5: Number of tweets influence on volatility as a dependent variable

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0005***	0.0001	0.003
LIBOR_ON	-0.0070***	0.0012	0.000
TWC_s	-0.0314***	0.0056	0.000
TWC_m	-0.0163***	0.0033	0.000
TWC_l		0 (omitted)	
cons	0.0715	0.0069	0.000
Wald chi2		44.41	
Sigma_u		0.0208	
Sigma_e		0.0383	
R-squared within		0.0678	
R-squared between		0.0344	
R-squared overall		0.0290	
Number of observations		41,745	
Number of groups		45	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

There were three dummies within a group created: twc_s (tweet count less than 100), twc_m (between 100 to 1000), and twc_l (over 1000). Small and medium number of tweets show significance in such a panel, though Boolean for tweets over 1000 was not included in regression due to the collinearity issue. Thus, there is a need to check each dummy group correlations separately. In order to do so, there was an isolated regression run. Each group had three different regression within 2019, 2020 and 2021.

At first, when regression was run within the whole timeframe and isolated within each group, it was observed that the least coefficient and thus influence on the volatility by LDT is observed within the group that is over 1000 tweets. And the highest beta observed is within the "less than 100 tweets" group (0.00024 vs 0.00114 corresponding betas). This result is quite expected as the deviation for the group of tweets that is less than 100 is way easier to spot for the user or trader that collects information from social media. Unfortunately, these results in isolation do not provide exhaustive information on the idea if there is any benchmark of tweets that should be noticed. To get closer to the clue there is decision to conduct test for each of group within each year – 2019, 2020 and 2021 for groups of small, medium and large number of tweets.

For the small group the results are the following: in 2019 it was the highest observable beta – 0.0011 and it was decreasing gradually within next two years. In the same time, the R2 of the regression as far as the significance of the LDT itself were also dropping. The number of observations dropped as well. Thus, it could be concluded that some companies that have a sound correlation in 2019 of LDT and volatility were just switching groups and moving to the medium group. Thus, it should be visible in those regression built.

For the observations with tweets that are counted less than 100 in 2019 and 2020 two regressions were estimated. The results are shown in Table 6 and Table 7 respectively.

Table 6: Small tweet count group influence on volatility as a dependent variable in 2019

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0012***	0.0001	0.000
LIBOR_ON	-0.0013	0.0042	0.753
cons	0.0323	0.0088	0.000
Wald chi2		100.20	
Sigma_u		0.0187	
Sigma_e		0.0203	
R-squared within		0.0197	
R-squared between		0.0162	
R-squared overall		0.0163	
Number of observations		4,459	
Number of groups		37	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Table 7: Small tweet count group influence on volatility as a dependent variable in 2020

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0008**	0.0003	0.024
LIBOR_ON	-0.0057***	0.0017	0.001
cons	0.0546	0.0057	0.000
Wald chi2		15.87	
Sigma_u		0.0313	
Sigma_e		0.0392	
R-squared within		0.0085	
R-squared between		0.0132	
R-squared overall		0.0016	
Number of observations		3,708	

Number of groups	34

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Regrettably, each regression within the medium group shows insignificance on the LDT variable. In the same time, within the large group it might be observed that since 2020 this factor is significant. The 2020 shows the highest beta with LDT coefficient – 0.0014 and it drops in 2021 to 0.00022. These fluctuations could be associated with increased activity in social media within the first COVID year. Furthermore, this year is bound to the increased subsidies and the huge amount of free time at home for lots of people within the developed world like EU and USA. All these factors have driven the retail investors activity and visibly enhanced the sensitivity of the investors to the posts in social media.

For the observations with tweets that are counted between 100 and 1000 in 2020 and 2021 two regressions were estimated. The results are shown in Table 8 and Table 9 respectively.

Table 8: Medium tweet count group influence on volatility as a dependent variable in 2020

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0002	0.0003	0.578
LIBOR_ON	0.0011	0.0024	0.657
cons	0.0579	0.0059	0.000
Wald chi2		0.39	
Sigma_u		0.0327	
Sigma_e		0.0418	
R-squared within		0.0004	
R-squared between		0.0489	
R-squared overall		0.0013	
Number of observations		4,465	
Number of groups		39	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Table 9: Medium tweet count group influence on volatility as a dependent variable in 2021

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0005	0.0003	0.222
LIBOR_ON	0.3111***	0.0813	0.000
cons	0.0247	0.0050	0.000
Wald chi2		14.64	
Sigma_u		0.0258	
Sigma_e		0.0335	
R-squared within		0.0069	
R-squared between		0.0457	
R-squared overall		0.0090	
Number of observations		5,654	
Number of groups		45	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

For the observations with tweets that are counted more than 1000 in 2020 and 2021 two regressions were estimated. The results are shown in Table 10 and Table 11 respectively.

Table 10: Large tweet count group influence on volatility as a dependent variable in 2020

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0014*	0.0007	0.053
LIBOR_ON	-0.0000	0.0015	0.959
cons	0.0699	0.0077	0.000
Wald chi2		4.30	
Sigma_u		0.0414	
Sigma_e		0.0413	
R-squared within		0.0077	
R-squared between		0.1352	
R-squared overall		0.0337	
Number of observations		1,682	
Number of groups		31	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Table 11: Large tweet count group influence on volatility as a dependent variable in 2021

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0002***	0.0001	0.000
LIBOR_ON	0.6561	0.4580	0.152
cons	0.0262	0.0325	0.420
Wald chi2		22.43	
Sigma_u		0.0413	
Sigma_e		0.0738	
R-squared within		0.0120	
R-squared between		0.0045	
R-squared overall		0.0088	
Number of observations		2,960	
Number of groups		41	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

These separate tests show that dummy for the group with the medium number of tweets (twc_m) is insignificant for this panel regression. Namely, deviation of tweet count within the medium tweet count group is insignificant and user literally doesn't notice it.

The reason why the medium number of tweets is insignificant to the regression perhaps might be explained as the banner blindness of the user/trader. On the one hand, this average number is being some kind of the regular "noise" to the reader of these short messages so it even doesn't bring any fluctuation to the market, but just keep it within the regular spread. On the other hand, the low number of tweets pushes the volatility down, probably because the retail investor just "forgets" about those stocks. The large amount, obviously, make the volatility in the asset to surge.

4.8. Meme group influence

In the beginning of the research within the dataset the "meme" dummy was tested and proved its significance. Thus, it is reasonable to test these group more thoroughly. For our research 2020 and 2021 are the most interesting period time frames as 2020 has suffered huge exogenous shock and 2021 was slightly more peaceful year while the world economy was adopting and incorporating effects of the shock of previous year.

Furthermore, in 2020 and 2021 the activity of retail investors has also surged mainly due to factors as free time and free money – a lot of subsidies were issued by western governments. A lot of these disposable assets (cash and time) were invested by retail investors on the market. Direct effect of this alternating environment could have been observed by the AMC and GME hype developed by Reddit community "wallstreetbets". That community was able to short squeeze several hedge funds who were in short position with stocks mentioned above.

Due to these reasons, exactly 2020 and 2021 are in the highest interest within this research. Thus, the dataset was tested in isolation within the meme group within 2020 and 2021 then out of the group within the same timeframe. The results of estimation for meme and non-meme companies for 2020 are presented in Table 12 and Table 13:

Table 12: Meme group influence on volatility as a dependent variable in 2020

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0006***	0.0002	0.001
LIBOR_ON	-0.0048	0.0036	0.181
cons	0.0761	0.0067	0.000
Wald chi2		12.54	
Sigma_u		0.0224	
Sigma_e		0.0588	
R-squared within		0.0037	
R-squared between		0.0100	
R-squared overall		0.0039	
Number of observations		4,092	
Number of groups		19	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Table 13: Non-meme group influence on volatility as a dependent variable in 2020

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0017***	0.0007	0.010

LIBOR_ON	-0.0011**	0.0005	0.021
cons	0.0328	0.0024	0.000
Wald chi2		11.23	
Sigma_u		0.0111	
Sigma_e		0.0228	
R-squared within		0.0176	
R-squared between		0.2223	
R-squared overall		0.0216	
Number of observations		5,763	
Number of groups		22	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Within both groups we see that beta for each group is significant, although there is quite a counterintuitive result – out of the meme group the magnitude of the effect on the volatility from the synthetic tweet variable (LDT) is 2.5 times higher: 0.16% vs 0.06%. Both numbers are quite low, though significant what provides the space for further research.

It is worth to mention that meme index ETF was introduced only in February 2021 by Roundhill Investments. But results within 2021 year still show that meme index companies are less influenced by LDT. The results of estimation for meme and non-meme companies for 2021 are presented in Table 14 and Table 15:

Table 14: Meme group influence on volatility as a dependent variable in 2021

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0003***	0.0001	0.001
LIBOR_ON	0.5883**	0.2727	0.031
cons	0.0320	0.0185	0.084
Wald chi2		17.40	
Sigma_u		0.0288	
Sigma_e		0.0669	
R-squared within		0.0113	
R-squared between		0.0002	
R-squared overall		0.0107	
Number of observations		5,588	
Number of groups		23	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Table 15: Non-meme group influence on volatility as a dependent variable in 2021

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0006***	0.0001	0.000
LIBOR_ON	0.1534***	0.0277	0.000
cons	0.0141	0.0018	0.000
Wald chi2		39.88	
Sigma_u		0.0120	
Sigma_e		0.0144	
R-squared within		0.0104	

R-squared between	0.0157
R-squared overall	0.0071
Number of observations	5,742
Number of groups	22

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

The year 2021 has the same sequence of volatility dependence on the tweet deviation. Within the group the beta is lower than out of the group. But for both groups these coefficients are lower: 0.057% vs 0.027%.

4.9. Finding shares that are sensitive to media coverage to create a portfolio/index

These results push the research further. Perhaps the group of assets chosen as meme stocks are not optimal in terms of such stocks that are influenced by social media coverage. Unequivocally, it is not clear what were the criteria for creating this ETF, though within our data there is a robust correlation, what is the reason to infer that the list of stocks in the ETF is not optimal even within our sample.

Additionally, there is an assumption that it is possible even within our sample to set the alternative ETF list that would be more sensitive to the social media coverage. In order to test this assumption, the sample was delimited by each single company to see if the company dummy shows significance. Then within the isolated data for each company simple OLS regression was run.

Within the panel regression it was discovered that each company has an individual significant dummy coefficient. The result of estimation is illustrated in Appendix 4. These results illustrate that each company within the observed period has its own specific correlation toward the social media coverage. Each company statistically scientifically differs from one another within the model.

After running simple OLS regression on each separate company it appeared that some regressions have significantly higher R2, thus LIBOR and LDT explain a lot of variance for those companies. The regressions with the R2 higher than 0.08 for the stocks were the following: AMD, BB, BBBY, BBY, CCL, CLF, COST, DWAC, GME, HOOD, TSLA.

With these 11 companies chosen above it was decided to create an alternative group named "ultra-meme". And then to see if this group will show higher sensitivity to the social media fluctuations. The result of estimation is presented in Table 16. In this sample group, there is a highest observed beta of 0.0024, that means that if LDT is changed by 1%, the volatility of the stocks of the companies under the sample group should also be adjusted by 0.24%. Considering that the model mostly incorporates the magnitude of the change of the tweets

(LDT) as an isolated source of information, it is a robust result with the solid confidence level that might be incorporated into the sophisticated financial models. In this group R-sq criteria also show high explanatory powers of the variables of the model.

Table 16: Ultra-meme group influence on volatility as a dependent variable

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0024***	0.0009	0.010
LIBOR_ON	-0.0143***	0.0041	0.000
cons	0.0625	0.0081	0.000
Wald chi2		13.85	
Sigma_u		0.0158	
Sigma_e		0.0457	
R-squared within		0.0938	
R-squared between		0.1038	
R-squared overall		0.0912	
Number of observations		11,555	
Number of groups		11	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Within the sample of these 11 companies the beta seems to be the highest if to compare to the meme and non-meme groups. The results of estimations for meme and non-meme groups are presented in Table 17 and Table 18. In meme group there is still quite a homogenous sample as the R-sq shows that the variance is explained by the variables is at the 5.6% within the cluster and 6.04% within the whole group. For the non-meme group the beta for LDT is higher, though the R-sq parameters are significantly lower. Furthermore, the sample count for non-meme group is more than two times higher than for the ultra-meme group and in meme group it is 1.5 times higher. The higher amount of observations could also explain higher Wald chi parameter in these groups.

Table 17: Meme group influence on volatility as a dependent variable

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0003***	0.0001	0.005
LIBOR_ON	-0.0155***	0.0035	0.000
cons	0.0788	0.0045	0.000
Wald chi2		24.69	
Sigma_u		0.0189	
Sigma_e		0.0559	
R-squared within		0.0567	
R-squared between		0.1363	
R-squared overall		0.0604	
Number of observations		17,581	
Number of groups		23	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Table 18: Non-meme group influence on volatility as a dependent variable

Factors	**Coefficients**	**Robust standard errors**	**P-value**
LDT	0.0010***	0.0003	0.000
LIBOR_ON	-0.0013***	0.0004	0.002
cons	0.0298	0.0025	0.000
Wald chi2		18.40	
Sigma_u		0.0010	
Sigma_e		0.0177	
R-squared within		0.0176	
R-squared between		0.3307	
R-squared overall		0.0193	
Number of observations		24,160	
Number of groups		22	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*

Such an outcome could be interpreted in the way that in accordance with our data sample we were unable to scientifically prove the idea of the composition of the meme stocks index ETF by Roundhill Investments. This might be either due to purely random composition of the index or due to the fact that the data of social media coverage collected by Bloomberg is slightly irrelevant.

Chapter 5: Conclusions and Further Research

5.1. Conclusions

The research conducted has shown the robust correlation between the social media coverage and the volatility. Unfortunately, based on the data used, it is not possible to validate the idea that social media coverage affects the returns of the assets. Thus, it is not possible to state any anomaly that would help to target returns. Hence, this work proves the correlation between generic factor of general tweet (LDT) and the volatility. Thus, it might be inferred that the increase in the coverage by social media of the asset pushes its risks forward with no observable potential sufficient robust returns. The higher the coverage, the more speculative the asset appears. So, we can conclude that *Hypothesis 1 "Tweets influence such derivative indices on share prices as return and volatility"* was proved partly.

Moreover, the research pays attention to the issue of tweet sentiment. It was expected that the negative tweets have stronger influence on the share price index then the positive ones, but the regression estimation showed no significance of the factors under consideration. *Hypothesis 2 "Media coverage influences share prices according to the behavioral finance principles that are reflected by the different investors' reaction on positive and negative tweets"* was rejected.

The next step of the research was to construct and test several derivative variables that can be taken into consideration. One such a variable generated is a lagged influence of tweets. It supposed to be a variable with a very short lag. So, the 1-day lag deviation 10-day moving average general tweet variable was used as the factor of the media coverage. It was significant in the econometric models tested. *Hypothesis 3 "Tweets influence share prices with a lag of a very short period, namely 1 day"* was proved.

Hypothesis 4 "The power of the media coverage influence depends on the features of companies" was proved partly. The power of the media coverage influence depends on the types of companies; they were divided into three types according to the potential attractiveness for investors who use social media. According to the econometric estimation for the sample and the period under research, the new and promising companies (so-called "meme") are more interesting and attractive to the investors then some other types.

Furthermore, it was found that the share price between 25 to 75 dollars is significant to the investors and rise of such shares can increase volatility because of trading activity. *Hypothesis 5 "The share price is crucial for a retail investor who has a restricted budget for their investments and is an active user of social media"* was proved partly.

In addition, the number of tweets has been validated as the significant parameter. Impressive results are shown to the group with the lowest and highest number of tweets. In the other words, if the median number of tweets is lower than 100 or higher than 1000, the influence of DLT on the volatility is way higher than at the medium group – from 100 to 1000 tweets daily coverage. This might be seemed as the user blindness – the usual pressure of the messages like a white noise – not noticed by the receiver. So, *Hypothesis 6 "The share price volatility depends on the frequency/number of tweets in social media"* was proved.

Finally, it was proved within the research, based on the data given, that the so-called meme index (ETF by Roundhill Investments) has no enough scientific validation. *Hypothesis 7 "The share price of the young companies from the list of ETF based on "the media coverage dependance" influences by tweets more than "blue chips"* was rejected. Alternative index, using the companies' stocks from our list, shows more sensitivity to the social media coverage.

5.2. Further Research

Current Thesis discloses the main idea that social media influences the market. Unfortunately, here we are limited with the resources and data to provide a thorough research on the behavior of the investors within the social media pressure. We are limited with the data provided by Bloomberg that obviously narrows the outcome.

In future, to provide more deliberate research it is valuable case to use proprietor research methods to analyze the sentiment of the tweets as far as the other social media sources. As an example, Reddit was not analyzed in the paper, though it is clear that it has a high influential power on the market (the r/wallstreetbets case of Jan. 2021).

In addition, there should be thorough analysis of the types of investors. As of today, the number of retail investors is growing and such companies as Robinhood and DeGiro provide swift access for the retail investors to the market that boosts the activity on the financial markets.

Furthermore, our research was limited to the daily data, while tweets and other social media messages tend to have a very short arm influence. People now have access to information 24/7 what leads to the acceleration of the transfer of information and thus shortens the correlation period. It might appear that especially short periods of minutes long after the announcements have the most correlation to the volatility as well as the returns.

And in the end, we hope that chiefly professional arbitrageurs could be the main beneficiaries of this research. Mostly these professional players value these tick-sized data alterations to be incorporated in their models to maximize profits.

References

Andersen, T.A., 1983, Price and Quantity Signals in Financial Markets, Journal of Economics, Vol. 43, No. 2, 109–129.

Ahern, K.R. and Sosyura, D., 2015, Rumor Has It: Sensationalism in Financial Media, The Review of Financial Studies, v. 28, n. 7, 2050–2093.

Ben-Rephael, A., Cookson, J. A. and Izhakian, Y., 2022, Do I Really Want to Hear the News? Public Information Arrival and Investor Beliefs (February 20, 2022). Available at SSRN: https://ssrn.com/abstract=3631760 or http://dx.doi.org/10.2139/ssrn.3631760

Boudoukh, J., Feldman, R., Kogan, S. and Richardson M., 2012, Which News Moves Stock Prices? A Textual Analysis, NBER Working Paper Series, working paper 18725, 46 p.

Bushman, R. and Pinto, J., 2019, The influence of short sellers on negative press coverage and price discovery, Kenan-Flagler Business School, University of North Carolina-Chapel Hill, 48 p.

Chen, C.-W., Pantzalis, Ch. and Park J.C., 2013, Press Coverage and Stock Price Deviation from Fundamental Value, The Journal of Financial Research, Vol. XXXVI, No. 2, 175–214.

Cookson, J.A. and Niessner, M., 2019, Why Don't We Agree? Evidence from a Social Network of Investors, The Journal of Finance, Volume 75, Issue 1, 173–228.

Dyck, A. and Zingales L., 2003, The Media and Asset Prices. Harvard Business School, University of Chicago, NBER&CEPR, 43 p.

Fama, E., 1970, Efficient Capital Markets. Review of Theory and Empirical Work, Journal of Finance, 25, 383–423.

Fama, E., 1991, Efficient Markets: II. Fiftieth Anniversary Invited Paper, Journal of Finance, 46, 1575–1617.

Frieberger, S.P., Lee, D., Puy, D., Ranciere, R. and Peria M.S.M., 2018, Media Sentiment and International Asset Prices, IMF paper, 1–22.

Gelman, S. and Sprenger G., 2014, What should be the Price of assets? Nobel Prizes in Economics 2013, Higher School of Economics Journal, No 1, 160–172.

Gervais, S., Kaniel, R. and Mingelgrin, D.H., 2001, The High-Volume Return Premium, The Journal of Finance, Vol. LVI, No. 3, 877–919.

Grossman, S. and Stiglitz, J., 1980, On the Impossibility of Informationally Efficient Markets. American Economic Review, 70, 393–408.

Haroon, O. and Rizvi, S.A.R., 2020, COVID-19: Media coverage and financial markets behavior – A sectoral inquiry, Journal of Behavioral and Experimental Finance, 27, 100343, 6 p.

Huang, T.-L., 2018, The puzzling media effect in the Chinese stock market, Pacific-Basin Finance Journal, 49, 129–146.

Jiao, P., Veiga, A. and Walter, A., 2020, Social media, news media and the stock market, Journal of Economic Behaviour and Organization, 176, 63–90.

Kahneman, D. and Tversky, A., 1979, Prospect Theory: An Analysis of Decision under Risk, Econometrica, Vol. 47, No. 2., 263–292.

Ke, Z., Kelly, B. T. and Xiu, D., 2021, Predicting Returns with Text Data (September 30, 2020). University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2019-69, Yale ICF Working Paper No. 2019-10, Chicago Booth Research Paper No. 20-37, Available at SSRN: https://ssrn.com/abstract=3389884 or http://dx.doi.org/10.2139/ssrn.3389884

Sewell, M., 2011, History of the Efficient Market Hypothesis, UCL Department of Computer Science, RN/11/04, 14 p.

Shiller, R.J., 1981, Do stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends? American Economic Review, 71, 421–436.

Su, Y., Qu, Y. and Yuxuan Kang, 2021, Online public opinion and asset prices: a literature review, Data Science in Finance and Economics, 1(1), 70–76.

Thaler, R.H. and Benartzi, S., 2004, Save More Tomorrow. Using Behavioral Economics to Increase Employee Savings, Journal of Political Economy, vol. 112, no. 1, pt. 2, 164–187.

Umar, Z., Adekoya, O.B., Oliyide, J.A., and Gubareva, M., 2021, Media sentiment and short stocks performance during a systemic crisis, International Review of Financial Analysis, 78.

Wong, J.-B. and Zhang, Q., 2021, Stock market reactions to adverse ESG disclosure via media channels. The British Accounting Review.

Yang, T., Liu, J., Ying, Q. and Yousaf, T., 2019, Media Coverage and Sustainable Stock Returns: Evidence from China, Sustainability, 18 p.

Appendix 1

Table: Literature review

Authors, year	Objects of research	Method	Data	Dependent variables	Results
Dyck A. and Zingales L., 2003	Stock prices which react to earnings announcement	Regression analysis	526 observations from the Wall Street Journal and other newspapers; GAAP and Street earnings, unexpected earnings, stock prices	Cumulative excess return; media report	Media spin affects the stock market response to earning announcement and this link is not benign, since media tend to report information biased in favor of companies.
Chen C.-W., Pantzalis C. and Park J.C., 2013	6,053 publicly traded U.S. firms	Panel data analysis	Number of news articles about every firm that appeared from 1995 to 2004 in the four major U.S. newspapers	Excess prices	Abnormal news coverage leads to mispricing that is rooted primarily in the fact that press coverage creates sentiment among investors
Ahern K.R. and Sosyura D., 2015	M&A deals and target firms asset prices	Panel data analysis	2,142 articles covering 501 rumors about 354 target firms in the time period 2000 – 2011	Cumulative abnormal returns in event time from 20 trading days before the rumor to 20 trading days after.	Investors do not fully account for the predictive power of merger rumors, leading to an initial target price overreaction and a subsequent reversal, consistent with limited attention.
Fraiberger S.P. et al., 2018	25 advanced and emerging countries	Construction a sentiment index; panel data analysis, VAR	More than 4.5 million Reuters articles in 1991 – 2015; daily equity flows from mutual funds investing in sixteen EMs in 2007 – 2015	Stock returns	Local news optimism attracts equity flows for a few days only, global sentiment optimism attracts them permanently
Huan T.-L., 2018	Shanghai and Shenzhen Main Broad 2598 common stocks	Cross-sectional return pattern, panel regressions	China Stock Market and Accounting Research database, 3,884,154 news stories from about 3442 news sources, 748 types of news. The sample is 2007–2014.	Stock returns	Even though highly controlled by the Chinese government, the financial media in China plays an important role in capital markets.
Bushman R. and Pinto J., 2019	Pilot Program firms	Cross-sectional analyses	Data from RavenPack News Analytics, which covers all news disseminated via Dow Jones Newswires and the Wall Street Journal	Three types of composite sentiment score	The increase in negative media tilt is significantly greater for pilot firms with lower media coverage intensity (number of articles), lower institutional ownership levels and higher bid-ask spreads
Yang T. et al., 2019	Chinese firms' stocks with high attention by	Panel data regression analysis	Monthly data of news reports from major Chinese newspapers, the "full-text database of China's major	Monthly raw stock returns, monthly DGTW	Media coverage has a more significant and positive influence on sustainable stock returns in the markets, dominated by

	individual investors		newspapers” provided by CmNeaKsIu, “Baidu index”	excess return	individual/immature investors
Haroon O. and Rizvi S.A.R., 2020	The benchmark indices for world and US; 23 sectoral indices for US from Dow Jones	EGARCH model	The Ravenpack finance for Panic Index, Global Sentiment Index and Media Coverage.	Return; volatility	Overwhelming panic generated by the news outlets are associated with increasing volatility in the equity markets
Jiao P., et al., 2020	Stocks mentioned in the news	Panel regression analysis	The Thompson Reuters MarketPsych Index (TRMI) database.	Stock volatility and turnover	Coverage by traditional news media predicts decreases in subsequent volatility and turnover, but coverage by social media predicts increases in volatility and turnover
Umar Z. et al., 2021	U.S. stocks of seven sectors covering consumer, energy, financials, healthcare, industrials, REITs, and technology.	TVP-VAR model	the Media Coverage Index (MCI); high short interest indices are sourced from Bloomberg in the period February 21, 2020 to June 04, 2021.	The 10-day volatility of the seven equity sectors	Healthcare and energy sector stocks behave as net recipients of both, returns and volatility; the MCI is more strongly connected with stock returns than with volatilities.
Wong J.B. and Zhan Q., 2021	US publicly traded companies.	Panel data analysis	A final sample of 331,517 observations, stock price data, the 4 factors from Fama, French and Carhart model (WRDS), S&P500 constituents, and the reputation risk (from the RepRisk database)	Firm value	Adverse ESG disclosure via media channels have a significant and negative impact on firm valuation.
Ben-Rephael A., Cookson J.A. and Izhakian Y., 2022	Stock data from the TAQ database, WRDS, CRSP, 8-K filings.	Panel data analysis	News coverage data from RavenPack Analytics Dow Jones News Wire; IBES data for earnings and analyst recommendation updates; CRSP data for trading volume, number of shares outstanding, and stock prices.	Uncertainty of belief, trading volume, volatility	Disagreement and trading are lower when the uncertainty of beliefs is higher.

Compiled by the author

Appendix 2

Table: List of the companies

Ticker	Company	Share prices, 10.1.2018 – 3.5.2021, $		Tweets per day, 10.1.2018 – 3.5.2021		(Non)meme
		min	**max**	**min**	**max**	
AAPL	Apple Inc.	34.419	182.707	0	14269	non-meme
AFRM	Affirm Holdings, Inc.	26.02	176.65	1	7840	meme
AMD	Advanced Micro Devices, Inc.	1.91	72.62	0	135305	non-meme
AMC	AMC Entertainment Holdings, Inc.	9.04	164.4599	0	9476	meme
AMZN	Amazon.com, Inc.	1237.23	3773.078	0	21612	non-meme
ATVI	Activision Blizzard, Inc.	38.8539	103.4116	0	10958	non-meme
BABA	Alibaba Group Holding Limited	73.28	319.32	0	6489	non-meme
BBBY	Bed Bath & Beyond Inc.	2.7	28.77	0	18006	meme
BB	BlackBerry Limited	3.43	53.9	0	3788	meme
BBY	Best Buy Co., Inc.	43.658	139.7632	0	1525	non-meme
BYND	Beyond Meat, Inc.	25	239.71	0	10371	meme
COST	Costco Wholesale Corporation	7.8	66.9173	0	2412	non-meme
CLF	Cleveland-Cliffs Inc.	2.5885	34.04	0	1912	meme
CCL	Carnival Corporation & plc	164.683	611.276	0	717	meme
DKNG	DraftKings Inc.	9.85	74.38	0	5220	meme
DWAC	Digital World Acquisition Corp.	12.62	175	485	57174	meme
EA	Electronic Arts Inc.	73.3474	150.1068	0	1823	non-meme
FB	Meta Platforms, Inc.	123.02	384.33	0	46434	non-meme
GME	GameStop Corp.	2.57	483	0	74211	meme
HOOD	Robinhood Markets, Inc.	9	85	68	33885	meme
JD	JD.com, Inc.	19.21	108.29	0	1712	non-meme
LCID	Lucid Group, Inc.	9.6	64.86	0	34365	meme
MCD	McDonald's Corporation	118.575	269.65	0	4344	non-meme
MRNA	Moderna, Inc.	11.54	497.49	0	6081	non-meme
MSFT	Microsoft Corporation	79.5194	348.9486	0	12728	non-meme
NFLX	Netflix, Inc.	185.6	700.989	0	16566	non-meme
NKE	NIKE, Inc.	59.0191	178.3716	0	4710	non-meme
NVDA	NVIDIA Corporation	30.9107	346.3688	0	8465	non-meme
PFE	Pfizer Inc.	24.4989	61.2443	0	17825	non-meme
PLTR	Palantir Technologies Inc.	7.25	45	44	8694	meme

PTON	Peloton Interactive, Inc.	17.21	171.09	0	10775	meme
RBLX	Roblox Corporation	29.52	141.5999	12	7816	meme
ROKU	Roku, Inc.	26.3001	490.7613	0	6127	meme
SAVA	Cassava Sciences, Inc.	0.76	146.16	0	6045	meme
SNDL	Sundial Growers Inc.	0.1381	13.22	0	40894	meme
SOFI	SoFi Technologies, Inc.	6.01	28.26	1	7980	meme
SPCE	Virgin Galactic Holdings, Inc.	6.7	62.8	0	15142	meme
SQ	Block, Inc.	32.33	289.23	0	3301	meme
T	AT&T Inc.	16.0165	25.2186	0	5955	non-meme
TLRY	Tilray Brands, Inc.	2.43	300	0	18875	meme
TSLA	Tesla, Inc.	35.398	1243.49	0	20846	non-meme
TTWO	Take-Two Interactive Software, Inc.	84.41	214.91	0	787	non-meme
WISH	ContextLogic Inc.	1.6	32.8499	5	20276	meme
WMT	Walmart Inc.	76.1697	160.77	0	5554	non-meme
ZM	Zoom Video Communications, Inc.	36	588.84	0	6085	non-meme

Appendix 3

Table: List of the variables

Variable in Stata	Explanation	Measure
VOLATILITY	Daily volatility of the price of the asset. Difference between the highest price and the lowest divided by the open price.	percent
TWC	Daily total tweet count for each company.	number
NTWC	Daily negative tweet count for each company.	number
PTWC	Daily positive tweet count for each company.	number
LIBOR_ON	Overnight LIBOR rate	percent
LDT	One day lagged daily deviation tweet count for each company of the 10 days moving average. Daily tweet count divided by the 10 days moving average of the t-1 day.	percent
HYPE	Dummy for the company that is under hype pressure.	Boolean
USER	Dummy for the company that's business based on the internet user activities.	Boolean
MEME	Dummy for the company of the meme index (Roundhill Investments ETF).	Boolean
TWC_s	Dummy for the daily tweet count that is less than 100.	Boolean
TWC_m	Dummy for the daily tweet count that is between 100 and 1000 tweet count.	Boolean
TWC_l	Dummy for the daily tweet count that is more than 1000.	Boolean
P_l	Dummy for the daily price of the asset that is less than 25 $ per share.	Boolean
P_m	Dummy for the daily price of the asset that is between 25 and 75 $ per share.	Boolean
P_h	Dummy for the daily price of the asset that is more than 75 $ per share.	Boolean

Appendix 4

Table. Companies' influence on volatility as a dependent variable

The table summarizes the econometric estimation results for the model with "volatility" as a dependent variable. Factors in the first columns are independent variables such as LDT – 1-day lag deviation of 10-day MA of general tweet count, LIBOR_ON – daily LIBOR, and tickers for all the companies of the sample. They are significant. F statistic shows that overall results are significant. The R-squared shows how much of the dependent variable variance is explained by the independent variables; it is 28 percent.

Factors	Coefficients	Robust standard errors	P-value
LDT	0.0004***	0.0001	0.000
LIBOR_ON	-0.0065***	0.0002	0.000
tick			
AFRM	0.0542***	0.0024	0.000
AMC	0.0554***	0.0021	0.000
AMD	0.0202***	0.0008	0.000
AMZN	0.0015**	0.0006	0.011
ATVI	0.0046***	0.0006	0.000
BABA	0.0057***	0.0006	0.000
BB	0.0222***	0.0011	0.000
BBBY	0.0379***	0.0014	0.000
BBY	0.0068***	0.0007	0.000
BYND	0.0352***	0.0015	0.000
CCL	0.0177***	0.0011	0.000
CLF	0.0293***	0.0009	0.000
COST	-0.0048***	0.0005	0.000
DKNG	0.0357***	0.0014	0.000
DWAC	0.0868***	0.0082	0.000
EA	0.0045***	0.0006	0.000
FB	0.0044***	0.0006	0.000
GME	0.0582***	0.0029	0.000
HOOD	0.0454***	0.0039	0.000
JD	0.0137***	0.0007	0.000
LCID	0.0422***	0.0030	0.000
MCD	-0.0056***	0.0006	0.000
MRNA	0.0436***	0.0013	0.000
MSFT	-0.0013**	0.0006	0.020
NFLX	0.0103***	0.0007	0.000
NKE	-0.0013**	0.0006	0.020
NVDA	0.0140***	0.0007	0.000
PFE	-0.0021***	0.0005	0.000
PLTR	0.0347***	0.0022	0.000
PTON	0.0388***	0.0013	0.000
PBLX	0.0389***	0.0019	0.000
ROKU	0.0343***	0.0009	0.000
SAVA	0.0746***	0.0028	0.000
SNDL	0.0842***	0.0037	0.000
SOFI	0.0391***	0.0027	0.000
SPCE	0.0288***	0.0016	0.000
SQ	0.0245***	0.0009	0.000
T	-0.0044***	0.0005	0.000
TLRY	0.0620***	0.0019	0.000
TSLA	0.0235***	0.0010	0.000
TTWO	0.0078***	0.0007	0.000
WISH	0.0546***	0.0023	0.000
WMT	-0.0052***	0.0023	0.000

ZM	0.0264***	0.0011	0.000
cons	0.2843	0.0005	0.000
F (46,41694)		422.99	
Prob > F		0.0000	
R-squared		0.2832	
Number of observations		41,471	

****, ** and * are statistical significances at 0.01, 0.05 and 0.10 level, respectively*