Assessing Social Media's Influence on Asset Price Synchronization: An Econometric Analysis of Private Information Impact

Abstract

This paper explores the impact of public information on the incorporation of private information into asset prices, particularly during periods of high price non-synchronization. Using time-series data from investing.com and Reddit, the study employs regression and correlation analyses to assess the predictive power of social media data on asset price movements. The findings suggest that while traditional market indexes failed to explain certain stock behaviours, social media-derived variables showed potential. This research contributes to understanding how non-traditional information sources influence asset pricing and systemic risk.

Keywords: Pirce non-synchronization, Asset Pricing, Social Media, Investment behaviour.

Introduction

This paper explores the relationship between price and the information it conveys in the market. We examine whether the price consistently reflects the true value of the asset or commodity in the market.

In the evolving landscape of financial markets, the dynamics of asset price formation and the interplay between public and private information have garnered significant attention. This study delves into the intricate relationship between asset prices and the information they reflect, exploring the extent to which prices in financial markets encapsulate the true value of assets and how information dissemination impacts this process. Amidst a plethora of studies examining market efficiency and information asymmetry, our research contributes a nuanced analysis of the influence of both public and private information on asset prices.

Central to our investigation are two pivotal research questions: First, we assess the measurable impact of public information dissemination on the incorporation of private information into asset prices. Second, we identify the key factors enhancing this transfer of information. Through these inquiries, we aim to uncover whether asset prices offer insights into systemic risks stemming from investor irrationality and individual decision-making processes.

Our empirical approach utilizes a rich dataset comprising information from investing.com and social media platforms like Reddit, adopting a multiple-linear regression framework to dissect the influences on asset prices. By analysing the nuances of price formation in the context of modern information flows, this study not only sheds light on theoretical aspects of market efficiency but also offers practical insights for investors and policymakers aiming to navigate the complexities of financial markets effectively.

The research hypotheses can be presented in following format.

$$High\ P_{Non-Synchronicity} = High\ Information_{Private}$$

$$Low\ P_{Non-Synchronicity} = Hight\ Information_{Public}$$

This means that in a moment when the asset price non-synchronization coefficient is higher, the explanatory power of private information for the variance of asset price is higher, than in a moment when the price non-synchronization coefficient is lower.

Both research questions could be tied into this one research hypothesis. The measurable impact is the price-non synchronization that can determine by quantitative means, and the private information is the variable that offers the factors and characteristics to explain this impact.

The market generates and disseminates information in transactions (Chen, Goldstein and Jiang, 2006; Glosten and Milgrom, 1985), categorized as public and private. Both have significant effects on market prices (Roll, 1984). The volume of available information has surged over the years, making it more accessible than ever before.

The impact of this information influx on price formation in asset markets remains underexplored. It necessitates deeper analysis to understand how the increasing data affects price formation, not solely driven by market-related factors. Prior research has primarily focused on detecting and estimating public information (price non-synchronicity), overlooking the influence of private opinion on asset classes. The influence of private opinion has received a new path to affect directly the asset price through social media. For this reason, it would be beneficial to address its effect.

For this determination, we use the price non-synchronization coefficient. It measures the variation in an asset's return that cannot be explained by market or industry returns (Durnev et al., 2003; Long et al., 2020). It is calculated by expressing the relationship between idiosyncratic volatility and systemic volatility (Chan et al., 2013; Long et al., 2020). To determine this coefficient, we use the definition of Morck, Yeung, and Yu from the year 2000. In their research, they used the R^2 of the capital asset pricing model (CAPM).

This paper aims to continue the work that Richard Roll started in 1984, by researching how this type of non-quantitative information can affect and transfer into the asset pricing. The field has been studied previously mainly from events and perspectives when companies' own management has taken advantage of using the information in their trading that has not been publicly available. For example, Baker and Wurgler's (2006) research. In our research, the definition of private information considers the information that private individuals are publishing on their social media accounts, and how this information can start to lead to the asset price when the information itself has transformed into prices through the individual's own transactions, that may have affected from the information on these social media platforms.

Literature Review

Asset Price Formation

One item affected by information in the market is the asset's price (Glosten and Milgrom, 1985; Chen et al., 2006). Previous research demonstrates the market's capacity to generate and transfer information (Roll, 1984; Chen et al., 2006). Richard Roll's 1984 study on weather forecasts' impact on orange juice futures prices revealed that forecasted cold weather influenced futures prices (Roll, 1984). While weather alone couldn't account for all price changes, it did contribute significantly to price volatility literature (Roll, 1984). Fama (1991) extends this discussion by emphasizing that prices reflect all available information, reinforcing the concept of market efficiency. Shiller (2003) introduces behavioural finance, suggesting that psychological factors and investor sentiment also play crucial roles in asset pricing. Tetlock (2007) examines how media content affects stock market prices, highlighting the role of information dissemination speed. Baker, Bloom, and Davis (2016) introduce an index to measure economic policy uncertainty, showing how policy-related information can influence markets. Engelberg, Sasseville, and Williams (2012) investigate the impact of television finance programs on stock prices, offering insights into specific types of media and their effects on investor behaviour.

Considering the research's age and technological progress since 1984, it's essential to acknowledge the increased accessibility and speed of information today. Additionally, orange juice futures represent a niche product from an investor's perspective. Despite these limitations, the research provides a valuable example of non-financial information's influence on asset prices. The inclusion of more contemporary studies from economic scientific journals adds depth to our understanding of how not only information but also investor sentiment, media influence, and policy uncertainty can significantly affect asset prices and market dynamics.

Multiple Linear Regression

Multi-linear regression is a statistical technique that models the relationship between a dependent variable and two or more independent variables. In econometrics, it is used to gain information on how multiple different factors could affect the economic outcome. This method predicts the dependent variable based on the independent variables, considering the possibility of interactions between them. Such an approach allows isolating the effect of each independent variable while controlling for the influence of others, providing a better understanding of the underlying factors of economic phenomena. The importance of multi-linear regression in econometrics stems from its ability to handle complex, real-world situations where multiple variables interact in determining an outcome (Greene, 2018; Gujarati & Porter, 2009).

Further enriching this discussion, Angrist and Pischke (2009) highlight the 'Credibility Revolution' in empirical economics, emphasizing improved research designs, including multi-linear regression, that enhance the reliability of econometric analyses. Wooldridge (2010) offers a comprehensive look at econometric analysis across cross-section and panel data, showcasing the versatility of multi-linear regression in economic research. Stock and Watson (2003) provide an accessible introduction to econometrics, including multi-linear regression, underscoring its foundational role in economic analyses. Lastly, Imbens and Wooldridge (2009) discuss recent developments in the econometrics of program evaluation, illustrating how multi-linear regression models are pivotal in assessing the impacts of economic programs and interventions.

These contributions from economic scientific journals underscore the evolving nature of multi-linear regression in econometrics, its critical application in analysing economic outcomes, and its significance in the broader context of empirical economic research.

Price non-synchronicity

The amount of information affecting asset prices significantly impacts investment decision-making and asset volatility. It is commonly assumed that asset prices reflect public information (such as annual reports and press releases) related to the asset (Chen et al., 2006; Long, Huaigang & Zaremba, 2020). Another type of information influencing asset price formation is private information, which is the focus of this paper. Private information becomes evident in asset prices following individual transactions (Chen et al., 2006; Roll, 1988). This type of information can lead to sudden and unexpected changes due to varying investment objectives and irrational decision-making. Measuring the amount of private information in asset prices involves examining price non-synchronicity (Chen et al., 2006; Long et al., 2020).

Price non-synchronicity measures the variation in an asset's return that cannot be explained by market or industry returns (Durnev et al., 2003; Long et al., 2020). It is calculated by expressing the relationship between idiosyncratic volatility and total or systemic volatility (Chan et al., 2013; Long et al., 2020). In 2000, Morck, Yeung, and Yu suggested that the R^2 coefficient from the Capital Asset Pricing Model (CAPM) could indicate asset synchronicity with the market (Morck et al., 2000; Long et al., 2020).

Expanding on this foundation, Jensen and Meckling (1976) introduced the concept of information asymmetry between managers and shareholders, highlighting how private information can significantly impact asset prices. Kyle (1985) provided a theoretical framework on how private information is incorporated into prices through trading, complementing empirical observations about price non-synchronicity. Shleifer and Vishny (1997) discussed how limits to arbitrage can lead to asset prices that do not fully reflect available information, including private information, thus contributing to price non-synchronicity. Baker and Wurgler (2006) examined how investor sentiment, potentially driven by private information, affects asset prices and leads to deviations from market efficiency. Lastly, Hong and Stein (1999) proposed a model where private information and investor behavior contribute to price momentum and reversals, offering insights into the mechanisms behind price non-synchronicity.

These contributions from economic scientific journals underscore the complex interplay between private information, investor sentiment, and market mechanisms in determining asset prices, providing a richer understanding of the factors influencing investment decision-making and asset volatility.

The equation 1 for the price non synchronization is following (Chen et al., 2006).

$$1 - R^2 \tag{1}$$

Where

R^2 – is the R-square of the regression presented in equation 2

$$R_{i,j,t} = \beta_{i,0} + \beta_{i,m} * r_{mt} + \beta_{ij} * r_{j,t} + \varepsilon_{it} \tag{2}$$

Where

$R_{i,j,t}$ – is the return of an asset I in industry j at moment t.

$\beta_{i,0}$ – is the intercept term representing the expected return of the asset that is not explained by the factors in the model

$\beta_{i,m}$ – is the sensitivity (or factor loading) of the asset i to the market factor m at time t

r_{mt} – is the market return at moment t

β_{ij} – is the sensitivity of the asset i to the specific factor j at time t

$r_{j,t}$ – is the return of industry j at moment t

ε_{it} – is the error term, representing the idiosyncratic return of the asset i at time t that is not explained by the model

Richard Roll (1988) was the first to conclude that price non-synchronicity and private information are correlated. Roll also stated that the information transfers into the prices in two different ways. Both ways are visible in his paper R2 is through public information that can be other information on the market or industry itself or newly published statistics on the performance of the asset (Roll, 1988, Chen et al., 2006, p.7). The second one is through private information. It means information on the individual's actions and perceptions in the market (Roll, 1988, Chen et al., 2006, p.7)

The debate surrounding the R^2 coefficient's ability to determine the amount of private information in asset market prices or noise trading is multifaceted, raising questions about market efficiency and investor irrationality (Long et al., 2020). Randall Morck, Bernard Yeung, and Wayne Yu's 2000 research highlighted a connection between lower R^2 and investor property rights, suggesting that in jurisdictions with weaker investment and ownership legislation, private information is less impactful on asset prices compared to markets with more developed legal frameworks (Morck et al., 2000). This implies that higher market non-synchronicity indicates less capitalized information from asset prices.

The relationship between investor irrationality and the R^2 coefficient of the Capital Asset Pricing Model has been extensively explored, with studies suggesting that R^2 is more closely linked to noise trading, psychological biases, and irrational market sentiment. This would imply that a lower R^2 coefficient might be expected in a market efficient from an information perspective, although empirical findings show inconsistencies with this view (Long et al., 2020).

Recent studies, such as those by Dong et al. (2020), have applied evolutionary gaming models to analyze price synchronicity, revealing that markets adapt to the private information provided by investors demonstrating above-average market performance. This adaptation suggests a potential for market control or alteration through private information (Dong et al., 2020).

Long (2020) studied the price non-synchronization phenomenon in Chinese stock market. Unlike evidence from the U.S. market, where price non-synchronicity is often linked positively to future returns, this study finds a negative relationship in the Chinese market. It examines over 2700 companies from 1998 to 2018, revealing that less synchronized stocks predict lower future returns. This effect is attributed to the low-idiosyncratic volatility anomaly, suggesting that once absolute idiosyncratic risk is accounted for, non-synchronicity becomes irrelevant to future returns.

Further enriching this discussion, Baker and Stein (2004) explore how market liquidity can reflect investor sentiment, affecting asset prices beyond traditional information efficiency paradigms. Daniel, Hirshleifer, and Subrahmanyam (1998) delve into psychological biases and their contribution to market inefficiencies, complementing the debate on R^2's significance. Fama and French (1992) and Shiller (1981) provide foundational insights into how factors beyond mere information efficiency, including investor behavior and market sentiment, influence stock returns and market volatility.

These contributions underscore the nuanced interplay between private information, investor sentiment, and market mechanisms in asset pricing, highlighting the complexity of determining market efficiency and the role of irrational behavior in financial markets.

Estimator Evaluation

To ensure the reliability of estimators in evaluating private information, several diagnostic tests are employed to meet the assumptions underlying regression analysis. The Variance Inflation Factor (VIF) is pivotal for detecting multicollinearity, where a VIF above 5-10 signals problematic levels, potentially compromising the regression coefficients' reliability (O'Brien, 2007; James et al., 2013). However, VIF's effectiveness may be limited in complex models, necessitating a nuanced understanding of multicollinearity (Mansfield and Helms, 1982; Fox, 1991).

The Durbin-Watson (DW) statistic serves to identify autocorrelation among regression residuals, essential for maintaining the independence assumption in OLS regression. A DW value around 2 indicates no autocorrelation, but interpreting this statistic becomes more challenging with complex models or multiple lags (Durbin and Watson, 1950; 1951; Kennedy, 2008; Cribari-Neto & Zeileis, 2010).

The Jarque-Bera (JB) test evaluates the normality of residuals, a critical assumption in many econometric models. Significant deviations from normality, as indicated by a high JB statistic, can undermine the efficiency of parameter estimates and the validity of inferences (Jarque and Bera, 1980; 1987; Bera and Jarque, 1981; Thadewald and Büning, 2007). However, the JB test's sensitivity to sample size and outliers highlights the need for careful interpretation and possibly alternative approaches in certain contexts.

Further enriching this framework, literature on regression diagnostics and alternative methodologies provides additional tools for addressing these challenges. Fox (1991) offers a comprehensive overview of regression diagnostics, including strategies for dealing with multicollinearity. Cribari-Neto and Zeileis (2010) discuss advanced regression techniques that account for autocorrelation, while Koenker and Hallock (2001) introduce quantile regression as a robust alternative to traditional methods, particularly in the presence of non-normal residuals. Baltagi and Li (1995) delve into autocorrelation tests in panel data, expanding the toolkit for analyzing time-series and panel data beyond the conventional Durbin-Watson approach.

By conducting these tests and considering the insights from recent literature, researchers can more accurately evaluate the assumptions for estimators, enhancing the reliability of conclusions drawn from econometric analyses of private information.

Research Design

The design integrates quantitative analysis through regression and correlation studies, utilizing data from investing.com and Reddit. The data encompasses variables like the frequency of specific emoji usage as a proxy for private information and daily stock price changes for selected companies, alongside market and industry indices for a comprehensive market view. The period of study spans from January 29, 2021, to June 30, 2021, allowing for detailed temporal analysis. This design is structured to test hypothesis on the relationship between information dissemination and asset price movements, employing various statistical tests to ensure the reliability and validity of the results, such as Variance Inflation Factor (VIF) for multicollinearity, Durbin-Watson for autocorrelation, and Jarque-Bera for normality of residuals. The goal is to unravel the intricate effects of information flows on market dynamics and individual stock behaviours, contributing to the broader understanding of market efficiency and investor behaviour.

Method of Research

The main descriptions of the data are presented in the below tables.

Vars	N	Mean	SD	Median	Trimmed	Mad	Min
Private_information	106	-0,86577	258,754	0	-0,05785	2,9652	-2241
BBY	106	0,000432	0,020459	0,00195	0,00122	0,017791	-0,0927
GME	106	0,015758	0,17968	-0,00225	0,005321	0,062566	-0,6
Market	106	0,001228	0,008368	0,001	0,001378	0,006968	-0,0245
Industry	106	0,00081	0,010799	0,00095	0,001166	0,008599	-0,0305

Source: created by the authors based on the calculations in R-Studio

Vars	Max	Range	Skew	Kurtosis	Se
Private_information	2190	4431	-0,28301	69,19772	21,19795
BBY	0,0469	0,1396	-0,83892	2,808928	0,001987
GME	1,0394	1,6394	1,89518	11,58327	0,017452
Market	0,0238	0,0483	-0,22269	0,646426	0,000813
Industry	0,0282	0,0587	-0,37492	0,657178	0,001049

Source: created by the authors based on the calculations in R-Studio

Private Information (private_information): Quantitative measure of private information, included for testing hypothesis about the impact of private information on prices. This variable is consisting of the number of daily appearances of the rocket emoji on the social media platform.

Company-Specific Data (BBY, GME): These columns can be used to analyse the impact of information on individual companies' stock performance. The variables are consisting of daily changes in the percentage of the stock's closing prices.

Market and Industry Data: Comparing company-specific data with broader market and industry trends can provide insights into whether the observed effects are company-specific or part of a wider market trend. Also, these variables would let me compare the effect of rocket emoji appearance to the effect that the industry (S&P 500 large and medium size retailer index) or market (S&P500 Index) towards the GME stock.

From this data we conduct regression and correlation analysis, to understand the relationships between market prices and information flow. The hypothesis could be structured around testing if there is a significant correlation between the volume of private information and price movements, or if the asset price reacts more towards the private information than the market or industry movements.

Micro evidence / Disaggregated data

Date Range and Frequency: The dataset starts from January 29, 2021, to June 30, 2021, covering 106 days. This granularity allows for a daily analysis of market dynamics.

Private Information (private_information): Significant variance from -2241 to 2190, indicating days with very high negative or positive variation in the use of a Rocket emoji. Slightly negative mean (-0.866), suggesting a slight decrease in the use of emoji over the period. This column can be used to examine the impact of private information on stock prices on a daily basis.

Company-Specific Stock Performance (BBY, GME): The dataset contains 106 data points for each company. Best Buy (BBY): Ranging from -9.27% to 4.69%, indicating considerable volatility. Mean approximately 0.04%, suggesting a relatively stable average performance over the period. GameStop (GME) varying from -60% to 103.94%, showing extreme volatility, possibly due to specific market events. Mean 1.58%, indicating a generally positive performance. These give an understanding of how individual stock prices react to private information daily.

Market and Industry Indices: Market index ranging from -2.45% to 2.38%, indicating daily market movement. Daily mean 0.12%, suggesting a slight overall positive trend in the market. The Industry Index varies from -3.05% to 2.82%, showing industry-specific fluctuations, with a mean of 0.08%, indicating a slight positive trend in the industry.

This dataset offers possibilities to conduct micro-level analysis of the impact of private information on specific stocks (BBY and GME) and compare these effects with broader market and industry trends. This is allowed using daily data, that gives detailed examination of the relationship between information flow and stock price movements, particularly in the context of market events that could cause extreme price fluctuations.

Macro Evidence / Aggregate data

Days with Positive vs. Negative Private Information: There are 68 days with positive private information and 66 days with negative private information, indicating the balanced distribution of positive and negative information days among the dataset.

Average Performance of Individual Stocks (BBY, GME): BBY has an average movement of approximately 0.043%, suggesting a relatively stable performance over the observed period, while GME has an average movement of 1.576, which indicates more volatile and positive trend compared to BBY.

Average Market and Industry Movements: The average market movement is 0.123%, indicating a slight overall positive trend in the market during the period. The average industry movement is 0.081%, also suggesting a marginal positive trend in the industry.

These macro-level insights highlight trends and averages, providing a general overview of market dynamics, the impact of private information, and the performance of specific stocks and sectors over the observed period. This is important for understanding the overall market sentiment and the general impact of information flow on market and industry trends.

Cross-section and time-series variance

Time-Series Variance: Private Information variable variance is 66,953.62, indicating a high level of fluctuation in private information over time. BBY has a variance of 0.000419, indicating relatively stable performance over time. GME has a higher variance of 0.032285, reflecting its more volatile performance. The market index variance of 0.000070 suggests a relatively stable market trend over the observed period. The industry index has a variance of 0.000117, the industry trend, which indicates stable progress.

Average Cross-Sectional Variance: The average cross-sectional variance for variables BBY, GME, market, and industry, is 0.0082.

Determining the data Panel / Cross-sectional: Because BBY and GME are the only entities considered, and the market and industry data are general indicators, providing the market and industry information, which are not tied to specific entities, the dataset cannot be qualified as panel data. By construction it is closer to a time series dataset focused on these two entities which has market and industry context. Also, the data won't pass the qualifications for pseudo-panel data, which means that it cannot be considered as that.

Nature of observed variables

Private Information: The variable is a quantitative measure of private information, represented by the number of daily appearances of the rocket emoji in social media discussions related to GameStop. Given that it's measured as a count (number of appearances), this variable is of the count type.

Company-Specific Data (BBY, GME): These variables represent the daily change in the stock prices of Best Buy (BBY) and GameStop (GME), measured in percentage. As they are expressed in percentage changes, they are continuous variables.

Market and Industry Data: These include the S&P 500 large and medium-sized retailer index and the S&P 500 Index. Similar to company-specific data, these are continuous variables, representing daily percentage changes in the indices.

Endogenous and exogeneous variables

The endogenous variables in my research are the GME and BBY stock prices, and the exogeneous variables are the market, industry indexes and private_information variable. This deviation is made based on equation 2 which was presented on page 3.

Proxies and measurement errors

Proxies: thinking of possible proxies, the 'private_information' variable, which measures the count of rocket emojis in social media discussions, acts as a proxy for the level of private information or interest in the market regarding specific stocks. It's not a direct measure of private information itself but is assumed to correlate with it. Also, variables BBY and GME can be thought of as proxies for company performance, or market conditions.

Measurement Error: The amount of measurement error can never be fully zero, but with certain ways and methods that have been also presented earlier in the text, their amount has been tried to minimize as low as possible. For example, through standardization of the data collection through API and using a standardized script, with strict limitations in counting the information from the social media data.

In the 'private_information' variable, measurement error could arise if the count of rocket emojis does not accurately reflect the true level of private information or sentiment. For instance, the use of emojis can be influenced by trends unrelated to market information, or some relevant discussions may not use emoji at all.

Stock prices BBY and GME can also be subject to measurement error. While they reflect the market's valuation of a company at a given time, this valuation can be influenced by short-term market volatility, speculative trading, or external events not related to the company's fundamental value. Although this valuation is the thing that is the thing that we want to measure and research in our study.

Unobserved Heterogeneity and Reverse Causality

Unobserved heterogeneity: This study acknowledges that there could be factors affecting BBY and GME stock prices that are not captured by the observed variables private_information, market, and industry indices. Some examples of these are internal company decisions and undisclosed financial information.

Reverse Causality: This research assumes that variables like private_information influence stock prices, it is also rational to note that changes in stock prices could affect the volume of social media discussions. This would imply a feedback loop where the dependent and independent variables influence each other causing endogeneity. This has been considered when looking at the information and specific words from the social media set. For example, words like buy, sell and hold have been left out from the notion, as from their context they could be seen to have a direct connection to affect the market behaviour, but in general this small rocket emoji on the other hand cannot be linked directly into investors marker activities directly based on the context. In my research project, one of the future steps is also the wider content analysis for this data, whose aim is to look for other variables, that would have the same logic.

To produce the price non-synchronization coefficient, the following regression following equation 3 has been formed from equation 2.

$$GME_{i,j,t} = \beta_{i,0} + \beta_{i,m} * market_{mt} + \beta_{ij} * industry_{j,t} + \varepsilon_{it} \tag{3}$$

Where

$GME_{i,j,t}$ – is the return of an asset I in industry j at moment t.

$\beta_{i,0}$ – is the intercept term representing the expected return of the asset that is not explained by the factors in the model

$\beta_{i,m}$ – is the sensitivity (or factor loading) of the asset i to the market factor m at time t

$market_{mt}$ – is the market return at moment t

β_{ij} – is the sensitivity of the asset i to the specific factor j at time t

$industry_{j,t}$ – is the return of industry j at moment t.

ε_{it} – is the error term, representing the idiosyncratic return of the GME at time t that is not explained by the model

In this case, the estimators for predicting the market movement of GME would be the market and industry.

To evaluate the effect of private information on the movement of a GME following equation 4 has been formed.

$$GME_{i,j,t} = \beta_{i,0} + \beta_{i,m} * private_{information_{mt}} + \varepsilon_{it} \tag{4}$$

Where

$GME_{i,j,t}$ – is the return of an asset I in industry j at moment t.

$\beta_{i,0}$ – is the intercept term representing the expected return of the asset that is not explained by the factors in the model

$\beta_{i,m}$ – is the sensitivity (or factor loading) of the asset i to the market factor m at time t

$private_{information_{mt}}$ – represents the private information at time t related to the market or some specific aspect of GME

ε_{it} – is the error term, representing the idiosyncratic return of the GME at time t that is not explained by the model

To handle the possible heteroskedasticity in the outcome following instrumental variables equations 5 and 6 are formed.

$$Private\ Information * = \pi^0 + \pi^1 Z + u \tag{5}$$

Where

$Private\ Information *$ – Represents the latent (unobserved or underlying) private information variable

π^0 – is the intercept term, representing the baseline level of private information when the instrumental variable Z is zero

π^1 – Measures the sensitivity or effect of the instrumental variable Z on the latent private information

Z – is the instrumental variable (or exogenous variable) in the model

u – is the error term, capturing the variation in private information that is not explained by the instrumental variable Z

$$GME = \alpha_0 + \alpha_1(private\ information) + \alpha_2\left(\frac{Market}{Industry}\right) + \varepsilon \tag{6}$$

Where

GME – is the predicted value of the dependent variable

α_0 – is the estimated intercept

$\alpha_1(private\ information)$ – is the estimated coefficient for the private_information variable

$\alpha_2\left(\frac{Market}{Industry}\right)$ – is the estimated coefficient for the market and industry indexes variable

ϵ – is the error term of the model.

In this case, the private_information variable as an estimator appears to be suitable to test the research hypothesis.

Regression and Correlation Analysis

Correlation

The following correlation table from the variables is being produced to check the relationship between variables.

	Private Information	BBY	GME	Market	Industry
Private Information	1	-0,07348	0,383274	-0,29395	-0,16371
BBY	-0,07348319	1	-0,044004	0,555193	0,578355
GME	0,38327422	-0,044	1	-0,10309	-0,13573
Market	-0,29395069	0,555193	-0,10309	1	0,754856
Industry	-0,16371304	0,578355	-0,135729	0,754856	1

Source: created by the authors based on the calculations in R-Studio

- Private Information and BBY have a correlation of -0.073, which implies a weak negative correlation.
- Private Information and GME have a correlation of 0.383, which has a moderate positive correlation.
- Private Information and Market have a correlation of -0.294, indicating a weak negative correlation.
- Private Information and Industry have a correlation of -0.164, which has a weak negative correlation.

- BBY and GME have a weak negative correlation of -0.044.
- BBY and Market have a moderate positive correlation of 0.552.
- BBY and Industry have a moderate positive correlation of 0.579.
- GME and Market have a weak negative correlation of -0.104.
- GME and Industry have a weak negative correlation of -0.136.
- Market and Industry have a strong positive correlation of 0.746.

Multicollinearity (Variance Inflation Factor)

Following table presents the Variance Inflation Factors (VIF) from the data.

Private Information	Industry	Market
1,104088	2,344737	2,497713

Source: created by the authors based on the calculations in R-Studio

- Private Information has a VIF of approximately 1.104, which indicates low multicollinearity. This is typically considered acceptable and indicates that this variable does not have a strong linear relationship with other independent variables.

The industry has a VIF of approximately 2.344, which indicates moderate multicollinearity. It can be stated to be within acceptable limits, and it indicates that the variable has a linear relationship with other independent variables.

The market has a VIF of approximately 2.498, which also indicates moderate multicollinearity. Similar to industry, and it suggests a linear relationship with other variables in the model.

All the VIF values from the model are below the common threshold of concern, which literature indicates 5 or 10 depending on the source. This indicates that the model does not have severe multicollinearity issues based on VIF.

Durbin Watson

Lag	Autocorrelation	Durbin Watson	P-Value
1	-0,01621857	2,023	0,61528

Source: created by the authors based on the calculations in R-Studio

Autocorrelation: The estimated value of the autocorrelation at lag 1 is -0.01621857. This number is very close to 0, suggesting there is no autocorrelation. The Durbin-Watson statistic is 2.023. The result is very close to 2, which implies that there is no significant autocorrelation in the regression model's residuals.

P-value: The p-value is 0.886, which is much higher than the common significance levels of 0.05 or 0.01. This high p-value indicates that there is not enough evidence to reject the null hypothesis of no autocorrelation.

The alternative hypothesis is that the autocorrelation coefficient is not equal to 0. Given the results, with a high p-value and a D-W statistic close to 2, it can be stated that no indication of significant autocorrelation in the data at lag 1. This suggests that the residuals from the model are random and independent from one another, which indicates a well-fitting model.

Jarque-Bera (JB)

From Jarque Bera test following outcomes were received.

Private Information
X-squared is extremely high at 10633, which is above typical critical values for any reasonable significance level. The p-value in the result is essentially 0, indicating it is extremely low. Given the low p-value and the high X-squared value, the null hypothesis can be rejected with high confidence. This means that the variable private_information does not come from a normal distribution and has significant skewness and/or kurtosis

Game Stop
A large value in the outcome indicates that the sample data does not follow a normal distribution. In this case, the value is 686.78, which is significantly large and suggests a deviation from normality. The p-value is extremely small (less than 0.00000000000000022), which is practically zero. This implies that the null hypothesis can be rejected with a very high degree of confidence.

Best Buy
The outcome suggests that the data deviates from normality. In this case, the value is 50.453, which indicates a significant deviation from normality. The degree of freedom is 2 because it is based on the estimates of skewness and kurtosis. The p-value is small (0.00000000001107), 0, which means the null hypothesis of normality can be rejected with high confidence. The test result shows that the Best Buy variable from the 'rocket' dataset is not normally distributed.

Market
A low value how that the data may not significantly deviate from normality. The value here is 3.167, which can be low, if compared to the other variables. The p-value here is 0.2053, which is greater than the common significance levels. Data does not show sufficient evidence to reject the null hypothesis. The data is normally distributed, as the p-value is above conventional thresholds for statistical significance.

Index
A low value implies that the data is more likely to be normally distributed. In this case, the value is 4.8904, which is relatively modest. The P-value of 0.08671 is above the common levels, suggesting that the evidence is not strong enough to reject the null hypothesis of normality.

Price non-synchronization

The following regression tables were produced by using the CAPM model to derive the price non-synchronization of GME and BBY stocks.

	Dependent variable:	
	GME	BBY
	(1)	(2)
index	-0.063	0.299^{*}
	(1.723)	(0.159)
industry	-1.864	0.629^{***}
	(1.208)	(0.112)
Constant	0.033^{**}	-0.0003
	(0.013)	(0.001)
Observations	209	209
R^2	0.020	0.291
Adjusted R^2	0.011	0.284
Residual Std. Error (df = 206)	0.189	0.017
F Statistic (df = 2; 206)	2.118 (p = 0.123)	42.188^{***} (p = 0.000)
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Source: created by the authors based on the calculations in R-Studio

By deducting the R^2 of both regressions from 1, we could see that the price non-synchronization coefficient of GME is 0,98 over the 209 observations. While BBY has a value of 0,709 from the same length of time. BBY applying this outcome to our hypothesis, the outcome would imply that the amount of private information affecting the GME stock during this period has been significantly larger than with the BBY.

It is also notable that the GME stock is negatively associated with the market and industry indexes, which would imply that during the monitored period it has progressed against, both market and industry, while the BBY has a positive association with the market and industry, which would imply that it has followed their progress more closely during the monitored period.

It can be noted that in the case of a GME the market and industry indexes were able to explain only 2% of the variance during the monitored period, while in the case of BBY 29% of the variation could be explained by the market and industry index movements.

In case of a BBY high F statistic of 42.188 with a p-value of 0.000 indicates that the model is highly statistically significant, while in case of GME F statistic of 2.118 with a p-value of 0.123 suggests that there is not enough evidence to conclude that the model is statistically significant at conventional levels, with market and industry variables.

Rocket Emoji / Private Information variable

To determine how the private information variable, could explain the movement of GME stock, the following regression was concluded.

	Dependent variable:
	GME
private_information	0.0002^{***}
	(0.0001)
Constant	0.016
	(0.016)
Observations	106
R^2	0.147
Adjusted R^2	0.139
Residual Std. Error	0.167 (df = 104)
F Statistic	17.908^{***} (df = 1; 104) (p = 0.0001)
Note:	*p<0.1; **p<0.05; ***p<0.01

Source: created by the authors based on the calculations in R-Studio

The coefficient for private information is 0.0002 and it is statistically significant. This implies a positive association between private information and the dependent variable GME. For each unit increase in private information, GME is expected to increase by 0.0002 units, holding all else constant. The R-squared value is 0.147, meaning that the model explains 14.7% of the variance GME. This is a relatively low value, suggesting that private information alone does not explain a large portion of the variability in GME. The F statistic is 17.908 with a p-value of 0.0001, which is highly significant. This suggests that the overall regression model is statistically significant and that private information is a significant predictor of GME.

This outcome would imply that in the case of GME the private information variable, which is the variance of rocket emoji in the dataset is better explaining the variance in GME, compared to the market and industry indexes.

Conclusions

In conclusion, it can be stated that in the case of the Game Stop (GME) the market and industry indexes could not explain the behaviour of the stock during the monitored time. Although with the comparative company Best Buy, the stock was following positively the market index movement. From this, we can derive that in this specific period, there is some external factor that is affecting the movement of a GME stock.

When looking at the direct answer to our research question number one: what is the measurable impact of the dissemination of public information on the incorporation of private information into asset prices? The answer is that with one variable derived from the social media dataset, the explanatory power of the model is 14,7% from the variance in the asset price. The finding itself has high statistical significance, which would imply that by deriving more variables from the social media dataset the model itself can be improved. Although this work would require deeper content analysis of the data, to determine the key variables and themes from the data.

Answer to the research question number two: what are the key factors that enhance this transfer? Our findings would suggest that in the case of social media the repetition of variables in the data, during a short period of time has affected the transformation of information into the prices. In our data this was visible in points when the amount of rocket emojis was high in the social media dataset it had a positive effect on the GME stock price also. However, as time passed the effect of this emoji on the GME stock price decreased, which is in line with the Long (2020) findings.

When testing our research hypothesis, we can denote that in the case of GME, the coefficient for price non-synchronization was higher (0,98), than with the BBY (0,709), our hypotheses would also imply that in this situation the amount of private information, in this case the amount of information tied into the asset in social media would be high also. Based on our findings with the GME stock price the null hypothesis can be rejected.

This movement we aimed to explain by an external qualitative factor, which in this case was the information from social media. This approach can be seen as a continuation of Roll's (1984) work that he started with his study of weather information and its effect on the orange juice derivatives. Our findings also align with his findings, that this type of non-financial information has an effect on the asset price, but it won’t be the only explanatory variable affecting the asset price movement.

Based on Dong's (2020) findings that the private information in this context would adapt to asset prices in cases when the asset is performing above the market index, we could state that GME fills this description also. When comparing the GME’s performance towards BBY price we could see higher variance in GME price, which movement the private information can explain more effectively than market and industry indexes.

The findings of Long (2020) could also be proven in our work, in his work the finding was that in the US market, the price non-synchronization can be connected to the positive movement in the market, as it was also in our case with GME.

The explanatory power of the rocket emoji itself was not high, but it would give indication to the authors, that by developing the model and adding more variables from the social media data, the estimator could be enhanced.

To do this more thorough analysis of the social media data needs to be done through content analysis, which aims to come up and bring up more of these variables, that may be used as instruments in the model in future.

The authors aim to do this analysis in the next article, and later to conclude and test these findings. The hypothesis of this paper was to show that if the price non-synchronization coefficient is high during the monitored time the explanatory power of the private information should be higher also, which my results are showing to be true.

Although the debate on the nature of private information is critical for this paper and authors acknowledge that its defiance must be much more precise than just contextual differences from trading activities.

References

Angrist, J.D. & Pischke, J.-S., 2009. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2), pp.3-30.

Baker, M. & Stein, J.C., 2004. Market Liquidity as a Sentiment Indicator. *Journal of Financial Markets*, 7(3), pp.271-299.

Baker, M. & Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4), pp.1645-1680.

Baker, S.R., Bloom, N. & Davis, S.J., 2016. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), pp.1593-1636.

Baltagi, B.H. & Li, Q., 1995. Testing AR(1) against MA(1) Disturbances in an Error Component Model. *Journal of Econometrics*, 68(1), pp.133-151.

Bera, A.K. & Jarque, C.M., 1981. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), pp.255-259.

Chan, K., Hameed, A., & Kang, W. (2013). Stock price synchronicity and liquidity. *Journal of Financial Markets*, 16(3), 414–438

Chen, Q., Goldstein, I., & Jiang, W. (2006). Price Informativeness and Investment Sensitivity to Stock Price. Fourteenth Annual Financial Economics & Accounting (FEA) Conference (Archive).

Cribari-Neto, F. & Zeileis, A., 2010. Beta Regression in R. *Journal of Statistical Software*, 34(2), pp.1-24.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A., 1998. Investor Psychology and Security Market Under- and Overreactions. *The Journal of Finance*, 53(6), pp.1839-1885.

Dong, et al., 2020. [Title of the Article]. *Journal Name*, [volume number]([issue number]), p.7.

Durbin, J. & Watson, G.S., 1950. Testing for Serial Correlation in Least Squares Regression: I. *Biometrika*, 37(3-4), pp.409-428.

Durbin, J. & Watson, G.S., 1951. Testing for Serial Correlation in Least Squares Regression: II. *Biometrika*, 38(1-2), pp.159-177.

Durnev, A., Morck, R., Yeung, B., & Zarowin, P. (2003). Does greater firm-specific return variation mean more or less informed stock pricing? *Journal of Accounting Research*, 41(5), 797–836.

Engelberg, J., Sasseville, C. & Williams, J., 2012. Market Madness? The Case of Mad Money. *Management Science*, 58(2), pp.351-364.

Fama, E.F., & French, K.R., 1992. The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 47(2), pp.427-465.

Fama, E.F., 1991. Efficient Capital Markets: II. *Journal of Finance*, 46(5), pp.1575-1617.

Fox, J., 1991. Regression Diagnostics: An Introduction. *Quantitative Applications in the Social Sciences*, 7(79), pp.1-79.

Glosten, L.R. & Milgrom, P.R., 1985. Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics*, 14(1), pp.71-100.

Greene, W.H., 2018. *Econometric Analysis*. 8th ed. New York: Pearson.

Gujarati, D.N. & Porter, D.C., 2009. *Basic Econometrics*. 5th ed. New York: McGraw-Hill.

Gujarati, D.N. & Porter, D.C., 2009. *Basic Econometrics*. 5th ed. New York: McGraw-Hill.

Hong, H. & Stein, J.C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6), pp.2143-2184.

Imbens, G.W. & Wooldridge, J.M., 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1), pp.5-86.

James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013. *An Introduction to Statistical Learning*. New York: Springer.

Jarque, C.M. & Bera, A.K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, 7(4), pp.313-318.

Jarque, C.M. & Bera, A.K., 1987. A Test for Normality of Observations and Regression Residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2), pp.163-172.

Jensen, M.C. & Meckling, W.H., 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), pp.305-360.

Kennedy, P., 2008. *A Guide to Econometrics*. 6th ed. Malden, MA: Blackwell Publishing.

Koenker, R. & Hallock, K.F., 2001. Quantile Regression. *Journal of Economic Perspectives*, 15(4), pp.143-156.

Kyle, A.S., 1985. Continuous auctions and insider trading. *Econometrica*, 53(6), pp.1315-1335.

Long, Huaigang & Zaremba, Adam & Jiang, Yuexiang. (2020). Price nonsynchronicity, idiosyncratic risk, and expected stock returns in China. *Economic Research-Ekonomska Istraživanja*. 33. 160-181. 10.1080/1331677X.2019.1710229.

Mansfield, E.R. & Helms, B.P., 1982. Detecting Multicollinearity. *The American Statistician*, 36(3a), pp.158-160.

Morck, Randall, Andrei Shleifer, and Robert Vishny, 1990, The stock market and investment: is the market a side-show?, *Brookings Papers on Economic Activity* 2, 157—215.

O'Brien, R.M., 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41(5), pp.673-690.

Roll, R. (1988) R2. *The Journal of Finance*, 43, 541-566.

Roll, R., 1984. Orange Juice and Weather. *The American Economic Review*, 74(5), pp.861-880.

Shiller, R.J., 1981. Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *The American Economic Review*, 71(3), pp.421-436.

Shiller, R.J., 2003. From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives*, 17(1), pp.83-104.

Shleifer, A. & Vishny, R.W., 1997. The limits of arbitrage. *The Journal of Finance*, 52(1), pp.35-55.

Stock, J.H. & Watson, M.W., 2003. Introduction to Econometrics. *Econometrica*, 71(1), pp.327-330.

Tetlock, P.C., 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3), pp.1139-1168.

Thadewald, T. & Büning, H., 2007. Jarque-Bera Test and its Competitors for Testing Normality – A Power Comparison. *Journal of Applied Statistics*, 34(1), pp.87-105.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. *Journal of Economic Literature*, 48(1), pp.106-107.

Wooldridge, J.M., 2013. *Introductory Econometrics: A Modern Approach*. 5th ed. Mason, OH: South-Western Cengage Learning.

Appendix 1 (Code)

```
data <- data.frame(BBY$Date)
data <- data %>%
  select("date" = BBY.Date)
data$date <- ymd(data$date)
data$BBY <- BBY$`Change %`
data$GME <- GME$`Change %`
data$index <- index$`Change %`
data$industry <- industry$`Change %`
data_info <- describe(rocket)

write_xlsx(data_info, "C:\\Users\\ssaar\\Desktop\\econometrics homework\\d
ata_info.xlsx")
```

```
modelGME <- lm(data$GME ~ data$index + data$industry)
summary(modelGME)

modelBBY <- lm(data$BBY ~ data$index + data$industry)
summary(modelBBY)

stargazer(modelGME,modelBBY, type = "html",style = "all2", out = "GME.html
")

modelrocket <- lm(rocket$GME ~ rocket$private_information)
summary(modelrocket)

stargazer(modelrocket, type = "html",style = "all2", out = "rocket.html")

iv_model<- ivreg(GME ~  private_information| industry + market, data = roc
ket)
summary(iv_model)

modelsummary(iv_model, metrics = "all")

cor(rocket[c("private_information", "BBY", "GME", "market", "industry")])
# check correlation between independent variables

vif(lm(data$GME ~ data$index + data$industry)) # calculate VIF values

vif(lm(GME ~  private_information + industry + market, data = rocket)) # c
alculate VIF values

durbinWatsonTest(lm(GME ~  private_information + industry + market, data =
rocket))

durbinWatsonTest(lm(data$GME ~ data$index + data$industry))

jarque.bera.test(rocket$private_information)

jarque.bera.test(rocket$BBY)

jarque.bera.test(rocket$GME)

jarque.bera.test(rocket$market)

jarque.bera.test(rocket$industry)
```