ECONOMICS AND CULTURE 22(1), 2025

DOI: 10.2478/jec-2025-0006

THE TECHNOLOGICAL BRIDGE: R PROGRAMMING'S UTILITY IN CONVERTING SOCIAL MEDIA DATA FOR QUANTITATIVE FINANCIAL ANALYSIS

Alexey Litvinenko[1], Saarinen Samuli[2], Anna Litvinenko[3]

[1] *University of Tartu, School of Economics and Business Administration, Tartu, Estonia, litvinenko@ut.ee, ORCID: orcid.org/0009-0002-3179-1514*

[2] *Estonian Business School, Tallinn, Estonia, s.saarinen@windowslive.com*

[3] *Tallinn University of Technology, Department of Business Administration, Tallinn, Estonia, anna.litvinenko@taltech.ee, ORCID: orcid.org/0009-0004-9528-9341*

Received 14 April 2024; accepted 4 June 2025

Abstract

Research purpose. This study explores whether R programming can transform unstructured qualitative social media data into a quantitative format suitable for econometric modelling. It specifically examines how elements such as text, emojis, and sentiment from Reddit and X (formerly Twitter) can be converted into variables for regression analysis. With the aim to enhance the predictive power of traditional financial models using alternative data sources, the paper outlines comprehensive guidelines with specific technical steps, from scripting an API to extracting data from Reddit and X, through cleaning and tokenising to incorporating the data into regression models using R programming. The study addresses the growing need in financial economics to incorporate alternative data streams by offering a structured, replicable process for transforming high-volume, unstructured online content into statistically valid variables, thereby bridging the gap between qualitative market sentiment and quantitative modelling.

Design / Methodology / Approach. Focusing on the methodology and R scripts, this research adopts a quantitative approach, transforming qualitative social media data into a format suitable for multiple linear and instrumental variable regression models to assess the effect of social media signals on asset prices, with GameStop (GME) and Best Buy (BBY) as case studies. The process ensures reproducibility and includes open-source code, enhancing transparency and applicability for both academic and professional financial data analysis contexts.

Findings. The findings demonstrate that qualitative social media data can be quantified for financial analysis. It was effectively extracted, cleaned, and used for regression analysis. Results show that traditional market indicators fail to explain GME's price shifts, while the frequency of rocket emojis (interpreted as speculative sentiment) was statistically significant. BBY's returns, however, aligned more closely with market and industry indices, suggesting a lower influence of private sentiment.

Originality / Value / Practical implications. The research provides a replicable method for integrating social media data into econometric models, contributing new tools for analysing market sentiment. By adapting classical financial models to modern data sources, the paper opens new directions for asset pricing research. The paper provides technical tools created in R for use in econometric analysis, useful both for academics and practitioners.

Keywords: R programming; econometric analysis; CAPM; price non-synchronization; social media data

JEL codes: C58; C87; G14

Introduction

This paper presents a method of obtaining social media information from two different sources using API and discusses how the data can be transformed by R into a format suitable for quantitative research. Currently, the methods of transforming social media content into an analysable format are limited both

* Corresponding author. E-mail: *litvinenko@ut.ee*

in the academic and practical world. Typically, the data collection and analysis were done primitively, which limited the size of the populations and samples analysed. Authors used to create a working script capable of extracting and analysing the collected data efficiently. Therefore, to tackle the problems of this process, the present research paper aims at presenting and testing a script capable of transforming social media data into an analysable format, eliminating the unnecessary limitation of population and sample size.

The research's following question helps to achieve the aim of the paper and contribute to solving a problem: Does the use of R programming enable the effective transformation of qualitative data into a quantitative format suitable for regression analysis?

Through the research hypothesis, we assume that the R script presented in this paper can transform the qualitative data that is in the format of text, images, emojis and video into a quantitative numerical format that could be analysed through Instrumental variable regression and other traditional multiple linear regression methods.

This paper is tailored for presenting the tools and methods of data extraction, cleaning, and use in multiple linear and Instrumental Variable regression models. The present research explores the features of the existing models and proposes a new method of asset price forecasting through the use of the Capital Asset Pricing Model (CAPM) and its R^2 coefficient to determine the asset's price non-synchronicity towards the market in moment x. This paper offers a new view of William F. Sharpe's (1964) and John Lintner's (1965) CAPM and its empirical applicability to modern social media data.

It is important to mention that despite being foundational in financial theory, CAPM has faced significant empirical scrutiny. Scholars have questioned its practical reliability, highlighting the persistent gap between theoretical assumptions and real-world applications. Building on this discourse, Morck et al. (2000) introduced the use of the CAPM's R^2 coefficient to measure asset price synchronicity, offering a novel lens for interpreting market dynamics. Subsequent studies, including Long et al. (2020), have extended this approach across various markets, revealing regional differences in how price non-synchronicity relates to future returns. These findings underscore the need for refined models that can incorporate market sentiment and firm-specific information, particularly from emerging data sources, such as social media.

With the present research, we expand the existing findings and unravel the characteristics of the asset price in a moment when the price non-synchronization is high, although the focus of this paper is to provide the method to obtain this information through R programming.

The article starts with a literature review in Section 2, continues with research design and research method under section 3 and is followed by research results in section 4. The conclusion in section 5 discusses the main findings, answers the research question and highlights the contribution.

Literature review

This section presents key theories and concepts underlying the R script, focusing on the theoretical foundations and empirical applications of quantitative methods in finance. It examines the evolution of the Capital Asset Pricing Model (CAPM) and regression analysis techniques, particularly in the context of modern data sources, such as social media. While such data offers new opportunities, it also challenges traditional financial theories. By reviewing past studies and critiques, this section sets the stage for investigating how R programming can transform qualitative social media data into a predictive tool for asset price movements during periods of high market volatility. Additionally, it highlights the historical significance, adaptability, and contemporary relevance of these financial models in an era of dynamic data-driven decision-making.

Investment Risk

Investment risk analysis could be considered a cornerstone of understanding the dynamics of financial markets and improving decision-making strategies. Among the new modern approaches, the incorporation of qualitative data from new-age media sources, like social media, represents a significant advancement in assessing asset price movements. This study contributes to the evolving discourse by

employing R programming to transform qualitative data into quantitative formats suitable for econometric modelling. Investment risk evaluation increasingly relies on robust statistical tools, such as Multiple Linear Regression (MLR) and Instrumental Variable Regression (IVR), to address complex relationships between market variables and mitigate issues like endogeneity.

The CAPM could be seen as the basis for financial risk assessment, linking expected returns to systematic risk through its beta coefficient. However, its empirical shortcomings, including the oversimplification of risk factors and the exclusion of market inefficiencies, have prompted researchers to seek alternative frameworks. For example, the concept of price non-synchronization measured by CAPM's R^2 coefficient provides a niche view of market behaviour, highlighting firm-specific risks that deviate from overall market trends. Studies revealed varying implications of non-synchronicity across global markets, emphasizing the need for contextual analyses in risk assessment (Long et al., 2020).

MLR offers a logical approach to CAPM by analysing the effects of multiple predictors on portfolio returns. Studies like Peng and Li (2018) demonstrate MLR's utility in identifying macroeconomic influences, such as GDP growth and inflation, on asset performance. However, MLR's reliance on linear relationships and assumptions of predictor independence limits its applicability in scenarios with complex interactions or endogenous variables.

To overcome these challenges, IVR has emerged as a powerful tool for estimating causal relationships in the presence of endogeneity. The two-stage regression approach of IVR allows researchers to isolate the impact of endogenous variables by leveraging external instruments (Lahouel et al., 2021). Despite its strengths, IVR's effectiveness is contingent on the selection of valid instruments, a process often debated in the literature (Lal et al., 2024). For instance, authors have highlighted the limitations of applying IVR to social media-derived variables due to challenges in establishing robust causal pathways.

The integration of social media data into investment risk analysis introduces a novel dimension, as it captures real-time market sentiment and private information dissemination. Tokenization of qualitative data, such as emojis and textual patterns, into quantitative formats, facilitates its incorporation into regression models. This approach enhances the predictive power of traditional models by accounting for non-transactional data's influence on market dynamics. While the empirical robustness of such methodologies requires further exploration, their potential to refine risk evaluation frameworks is evident.

By bridging the gap between theoretical models and practical applications, this study points out the importance of adapting classical econometric tools to modern data sources. The findings not only validate the utility of R programming in transforming social media data but also establish the grounds for more sophisticated analyses of investment risk, incorporating a broader spectrum of variables to achieve more accurate and actionable insights.

Capital Asset Pricing Model

The Capital Asset Pricing Model (CAPM), represented in Equation 1.0, can be perceived as the foundation for explaining the relationship between systematic risk and the expected return on assets. According to CAPM, the expected return on an asset or portfolio is the sum of the risk-free rate and a risk premium, given by the asset's beta. Beta quantifies how sensitive the asset's returns are to fluctuations in the market. This model was first introduced by Sharpe (1964) in his seminal paper and later expanded by Lintner (1965), who reached similar conclusions. Together, these researchers proved that investors are compensated for the time value of money (reflected in the risk-free rate) and the level of risk assumed (reflected in the risk premium), with higher levels of risk corresponding to higher expected returns.

The CAPM model is widely recognized for its straightforward approach, making it a valuable tool for estimating the cost of equity and managing portfolios. However, empirical research (Black et al., 1972; Roll, 1977) has revealed gaps between its theoretical predictions and real-world outcomes, raising concerns about its practical reliability. This criticism highlights the importance of critically evaluating the CAPM's reliability in practical financial decision-making, emphasizing the need for a deeper understanding of its limitations and the specific conditions under which it remains valid or falls short.

The Capital Asset Pricing Model (CAPM) is often criticised for its empirical and theoretical limitations, including its failure to predict asset returns (Fama & French, 2004), the unobservability of the true market portfolio (Roll, 1977), and its unrealistic assumptions about investor behaviour and market conditions (Merton, 1973). These critiques have led to the development of multifactor alternatives, such as the Fama-French three- and five-factor models (Fama & French, 1993; 2015) and the Carhart four-factor model (Carhart, 1997), which aim to better capture the complexities of market risk.

Another limitation is known as price desynchronization. Price desynchronization, quantified by the R^2 from the CAPM, reflects how closely individual asset prices move with the market, with lower R^2 indicating greater idiosyncratic volatility (Morck et al., 2000). Research shows that price synchronicity varies by market conditions (Chordia et al., 2008; Hou et al., 2015), and in markets like China, lower synchronicity has been linked to lower future returns (Long et al., 2020). Besides criticism, the CAPM model has been rethought, analysed and developed by researchers and practitioners.

Andre F. Perold (2004) highlights the historical development and foundational role of the CAPM in linking risk and expected return, a view supported by Dempsey (2013), who underscore its continued relevance in finance education and practice. Extensions of the model have emphasized liquidity as a critical factor influencing asset pricing (Amihud & Mendelson, 1986; Pastor & Stambaugh, 2003; Acharya & Pedersen, 2005). At the same time, there are empirical limitations due to market portfolio proxies, while Shafer and Vovk (2019) propose a game-theoretic approach to derive CAPM without relying on investor preference assumptions. These debates and innovations reflect both the enduring influence and the evolving critiques of CAPM, prompting the development of more robust models to address its theoretical and empirical gaps (Bhattacharya, 2015). While the CAPM remains a fundamental framework in financial analysis and investment decision-making, its theoretical limitations and empirical shortcomings have led to ongoing refinements and the development of alternative models. The critical evaluation of CAPM underscores the need for more comprehensive approaches that account for market inefficiencies, firm-specific risks, and the growing role of alternative data sources in asset pricing.

Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical method that is derived from simple linear regression. It uses two or more independent variables to predict a dependent variable. This technique models the linear relationship between the dependent variable and multiple predictors, which allows the analysis of how various factors concurrently affect the outcome (Lee & Eid Junior, 2018). MLR is used for evaluating the strength of the association between the dependent variable and each independent variable while taking into account the influence of other predictors. Key components of MLR outcome are the multiple correlation coefficient, adjusted R-squared (which indicates the proportion of variance in the dependent variable explained by the model), and the interpretation of regression coefficients, including those for categorical and dummy variables (Hayes & Matthes, 2009). Additionally, MLR can show interaction effects, where the impact of one independent variable on the dependent variable varies depending on the level of another predictor (Rencher & Christensen, 2012).

In the field of investment risk, MLR has been applied to analyse the sensitivity of portfolio returns to multiple macroeconomic factors, examining the influence of market performance, real GDP, inflation, and unemployment on portfolio returns using MLR, demonstrating its utility in assessing investment risk (Peng & Li, 2018). Other researchers used MLR to predict stock market volatility, underlining its relevance in financial risk assessment (Smith et al., 2018). These examples underscore the applicability of MLR in evaluating how various factors impact investment outcomes, making it a valuable tool in financial analysis.

IV Model

Instrumental Variables Regression (IVR) is a two-stage method designed to address endogeneity and improve causal inference in financial models, particularly when conventional regressions suffer from omitted variable bias or simultaneity (Grace, 2021; Maydeu-Olivares et al., 2020). Despite its strength, IVR's effectiveness depends heavily on the quality of instruments, which often lack consensus and reduce robustness across studies (Saarinen & Litvinenko, in press). This was evident in Saarinen and

Litvinenko's research, where IV models were excluded from the final analysis due to their limited explanatory power when using social media-derived variables for asset pricing. Nonetheless, IVR remains a valuable tool in both academic and applied finance, especially in complex systems where causal clarity is crucial.

Research methodology

In this chapter, the research design and method are presented, and the theoretical as well as mathematical models behind the R script are introduced.

Research Design

The research uses a quantitative methodology focused on transforming qualitative social media data into a format that can be applied for regression analysis. This study aims to explore how R programming can be utilized to convert unstructured social media content, specifically from platforms like Reddit and X (formerly Twitter), into structured data that can be analysed using multiple linear and instrumental variables regression. Data is extracted using API scripts, cleaned, tokenized, and then modelled using R programming. The primary variables of interest include social media indicators (such as the frequency of specific emojis) and their relationship with asset price movements. By employing regression techniques through R programming, this research investigates the role of price non-synchronization in asset prices and the predictive power of social media data during high-variance market periods through R programming.

Since this research aims to transfer these models into R-Script and to test them, the process and logic behind the R-script are presented in this paper, providing a replicable methodology to apply similar techniques in both academic and business contexts. This design is suitable for expanding the empirical applicability of the CAPM in modern contexts where social media data plays a significant role in influencing market behaviour.

Research Method

To produce the price non-synchronization coefficient, the following regression formula (Eq. 1) was formed.

$$GME_{i,j,t} = \beta_{i,0} + \beta_{i,m} * Market_{mt} * Industry_{j,t} + \varepsilon_{it} \quad (1)$$

where

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

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

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

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, the following formula (Eq. 2) has been formed.

$$GME_{i,j,t} = \beta_{i,0} + \beta_{i,m} * Private\ Information_{mt} + \varepsilon_{it} \quad (2)$$

where

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

$Private\ Information_{mt}$ – is the number of rocket emojis presented per day on Twitter and Reddit platforms.

In this case, the estimator for predicting the market movement of GME would be the private information variable. The first stage of the IV model shown in the formula below (Eq. 3) is built in the following way.

$$Private\ Information = \Pi Market + \Gamma Industry + \epsilon \quad (3)$$

where

$Private\ Information$ – Daily variance of the rocket emoji in the dataset

$\Pi Market$ – Coefficient for market index

$\Gamma Industry$ – Coefficient for industry index

ϵ – is the error term

The second stage shown in the formula (Eq. 4) is constructed in the following way

$$GME = \beta Private\ Information\hat{} + \Delta Market + \Delta Industry + u \quad (4)$$

where

GME – Is the daily variance of GameStop stock

$\beta Private\ Information\hat{}$ – The coefficient for the effect of private information on the GameStop stock

$\Delta Market$ – Coefficient for the effect of Market index on the GameStop stock

$\Delta Industry$ – Coefficient for the effect of Industry index on the Game Stop stock

u – is the error term

The full script is available on GitHub. The direct link can be found in Saarinen (2023).

Research results

The process for the data collection started by forming the script for the API to extract the data from the service provider. In this case, the service provider is X and Reddit. The API was designed following the method outlined in lines 15 - 88 of the Data Extraction file provided via link (Saarinen, 2023). The API code was written in Python, as it is the programming language supported by both service providers for API functionality.

The first rows of code are dedicated to inputting our unique API key and password, which was obtained from the customer service of Twitter (now rebranded as X) in Autumn 2024. The following line specifies the Access token and its corresponding password for the API, which were also obtained from the service provider. The next line tests the functionality of the API. When the connection is successfully established, the script provides a confirmation message. The last section of the code handles the actual retrieval of the tweets using the api.user_timeline command. In the screen_name parameter, a list of usernames was provided for retrieving their tweets. In our example script, the number of tweets is limited to 10 to prevent extended running time in case the script is accidentally executed, ensuring its completion in just a few minutes.

For Reddit, the retrieval script can be found in lines 20 - 40 in the Data Extraction file. It follows similar logic in the Client_ID and Client_secret fields, where API keys should be inserted when obtained from the service provider, and the user_agent field is served for the e-mail that is registered for these credentials.

The next step was to create the hot_posts variable. The placeholder “Python” was replaced with the name of the subreddit which is aimed to be extracted. In our case, it is “WallStreetBets”. Additionally, in this example, the search was limited to the first 10 rows to prevent prolonged runtime in case the script was accidentally executed.

Following successful data extraction, a data-cleaning process and analysis were initiated. The first step was to download the necessary packages for these tasks. These processes are detailed in lines 91 – 109 of the Data Extraction file (Saarinen, 2023). The package reader was chosen to upload the data into the script from the CSV files that contain the stock prices, as well as market and industry index information.

Package varhandle was chosen for the cleaning of the data, especially its gsub function, which was useful in situations where it was needed to alter the decimal nominator of the values in rows by changing them from commas to dots. Package dplyr was used for the data cleaning because of its wide range of functions, like select, mutate, and filter. The package lubridate helped to alter the character-based data

values into date-type values. The tokenization of the data from Twitter and Reddit was done through the package quanteda.

Package readtext was used to upload data from the CSV file assisting in making a comparison between the read_csv function to see if these two functions have any difference in the outcome. The package ggplot2 was used to format the outcome of the analysis into a graphical format, while package quanteda.textstats helped to formalize the statistics from the tokenized text data.

To format the multiple linear regression and IV model tables into their publishing format, the Stargazer package was used. Package lmtest was chosen for the multiple linear regression analysis. Further, package IV reg was used for the creation of IV model regressions. The data was uploaded into the script by using the read_csv function, chosen for the csv files.

The first step in the data-cleaning process was to ensure that the time variable in the Reddit table was in the correct format. This process is described in lines 111 – 140 of the Data Extraction file (Saarinen, 2023). It involves converting the timestamp variable using the ymd_hms function from the lubridate package.

Next, we created a corpus from the same data using the corpus function from the quanteda package. This function essentially indexes the individual characters and words within the dataset. The summary function can then be used to inspect the outcome. The next step was to index the data. The token was formed from the rocket emoji, and the search for this token was done from the indexed data, using the tokens_keep and dfm function from the quanteda package.

Further, we calculated the daily frequency of the rocket emoji from this data using the textstat_frequency function from the quanteda.textstats package. The search was performed on the corp_red_dfm variable, which contains the timestamps and the count of the rocket emojis observed at specific times in the subreddit. In the example, the counting length was limited to the first 5 days for simplicity, but in research, it was conducted over a period exceeding 100 days. Finally, the results were grouped by the dates specified in the timestamp column.

After completing the grouping, we ensured that the timestamp column, renamed as “group”, was converted to date format rather than character format. This is achieved using the as.date function. Once this was done, we created the list object from the result, showing the total daily count of the rocket emoji. In the final step, with these daily totals, we calculated the differences between consecutive dates using the diff function. The result is saved in a variable named list.change.

Since the list of daily changes has a length of 149 instead of 150, we used the following code to remove any NA variables from the list, ensuring its length is adjusted to 149.

Further, we created the list.change2 variable, which represents the daily percentage change. This was calculated by dividing the change values in list.change1 by the corresponding values in the list.changex the variable and multiply the outcome by 100. In the final step, we removed all the NA values from the resulting data by the na.omit function. After this, we mutated the column name to a clearer format. This was done by using the mutate function from the dplyr package.

In the final step, the data on rocket emoji usage was exported to an Excel file for further analysis. This process is detailed in lines 138 and 139 of the Data Extraction file (Saarinen, 2023). With the social media variables successfully extracted and cleaned for analysis, the next step was to apply the same process to the stock price, market, and industry index variables. This step is relatively straightforward, as these variables are already formatted to reflect daily changes.

First, we converted the variables from character type to factor type. This allowed us to replace commas with dots and remove the % sign from the end of the values. This transformation is performed using the unfactor and gsub functions from the varhandle package. To finalize this process, we converted the variables to numeric type and divided the resulting values by 100 to format them as decimals for further analysis. These steps are detailed in lines 160-191 of the Data Extraction file (Saarinen, 2023).

From the cleaned stock price and index data, we created a new table named “data”. First, we extracted the “date” column from the BBY table and saved it as the date column in the “data” table. Next, we

added the percentage change columns from other tables to the "data" table, assigning each column its name. Finally, we used the "describe" function from the psych package to generate summary statistics for this table. This process is outlined in lines 76-78 of the Regression file (Saarinen, 2023).

Proceeding with the analysis, the first step was to create the regression models for IV and multiple linear regression. The regression models are visible in code lines 83-97 of the Regression file (Saarinen, 2023).

The multiple regression models were built using the lm function, and their results were summarized with the summary function in base R. The output was then formatted and finalized using the stargazer function from the stargazer package. Similarly, the IV model was constructed using the ivreg function, with its results summarized using the modelsummary function.

The final step of our analysis involved assessing the correlation among independent variables and calculating their Variance Inflation Factor (VIF) values to evaluate their suitability for inclusion in the model. This was performed using the base R functions cor and vif. Additionally, the Jarque-Bera test was applied to assess the robustness of the variable. These steps are detailed in lines 99-116 of the Regression file (Saarinen, 2023). The outcome of an Instrumental Variable regression table is presented below in Table 1.

Table 1. The results of an IV regression (Source: created by the authors)

Dependent variable:	**Intercept**	**Index**	**BBY**	**Summary Statistics:**
				Residual Std. Error: 0.1897
Estimate	0.03346	-2.32660	0.61397	Multiple R-Squared: 0.01245
Std. Error	0.01320	1.45640	0.70436	Adjusted R-Squared: 0.002861
t value	2.535	-1.597	0.872	Wald Test (2 df): 1.298, p-value: 0.2752
p-value	0.012 *	0.112	0.384	

Table 1 shows the result of an IV regression where the GME is the dependent variable. The results show that the coefficient for the index is -2.32660, but this is not statistically significant (p = 0.112). Similarly, the coefficient for BBY is positive (0.61397) but also lacks significance (p = 0.384). The intercept is statistically significant (p = 0.012), indicating a meaningful constant baseline effect in the model. The adjusted R^2 is minimal (0.0029), indicating that the model explains a very small portion of the variance in GME. Table 2 shows a regression table from the multiple linear regression detailed view by presenting results for both GME and BBY.

Table 2. Results of the MLR for GME and BBY (Source: created by the authors)

Variable	**GME (1)**	**BBY (2)**
index	-0.063 (1.723)	0.299* (0.159)
industry	-1.864 (1.208)	0.629*** (0.112)
Constant	0.033** (0.013)	-0.0003 (0.001)
Observations	209	209
R^2	0.020	0.291
Adjusted R^2	0.011	0.284
Residual Std. Error	0.189 (df = 206)	0.017 (df = 206)
F Statistic (p-value)	2.118 (p = 0.123)	42.188*** (p = 0.000)

As is visible from Table 2, for GME, the index coefficient (-0.063) is negative and not significant (p = 0.123). Similarly, the coefficient for the industry is also negative (-1.864) and lacks significance. The intercept (0.033) is significant ($p < 0.05$), aligning with Table 1's findings. For BBY, the results are more promising. The index coefficient (0.299) is positive and significant at the 10% level (p = 0.059), while the industry coefficient (0.629) is highly significant ($p < 0.01$). The intercept is insignificant. The R^2 for BBY (0.291) and its adjusted R^2 (0.284) indicate the model explains about 29% of the variance, which is higher than for GME. The results show significant differences between GME and BBY. BBY has stronger relationships with its predictors, especially the industry index, while GME's predictors don't have statistical significance. This difference highlights the variability in how the independent variables influence the dependent outcomes, with BBY being more predictable and affected by the variables tested. This could suggest that external factors, such as market trends, play a more significant role in explaining BBY's variance than GME's. From this outcome, it can be noted that in the case of GME, both IV regression and multiple linear regression with market and industry index as independent variables would indicate that in the case of GME, there is something else that is moving the asset price in the market at a given moment. To determine the predictive power of the Rocket emoji's variance in the Reddit dataset, the following Table 3 could be produced as the outcome with the script.

Table 3. A regression table where the number of rocket emojis per day was used as the independent variable. Source: created by the authors.

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 (p-value)	17.908*** (p = 0.0001)

The regression analysis presented in Table 3 shows a significant positive relationship between private_information and GME, with a coefficient of 0.0002 ($p < 0.01$), indicating that private information strongly influences GME's performance. The constant term, 0.016, is not statistically significant, suggesting that the baseline value for GME is not reliably different from zero in the absence of private information. The model explains approximately 14.7% of the variation in GME (R^2=0.147 R^2= 0.147 R^2=0.147), which is remarkable for financial data but also indicates the presence of other influential factors not included in the model. The residual standard error of 0.167 reflects the average deviation of observed values from predictions, while the F-statistic of 17.908 ($p < 0.01$) confirms the overall significance of the model. These findings emphasize the crucial role of private information in predicting GME, but the low R^2 suggests that additional variables are needed for a more comprehensive explanation of GME's movements.

Conclusions

From the analysis presented above, it is evident that the quantitative examination of social media data can be effectively achieved by the tokenization of qualitative data. This approach enables the processing and analysis of large datasets, thereby generating more generalizable insights derived from individuals' reflections shared on social media platforms.

In response to the research question: Does the use of R programming enable the effective transformation of qualitative data into a quantitative format suitable for regression analysis? The answer is affirmative. The findings of this study demonstrate that the technical approach discussed here effectively generates

data that can be employed in models with a high degree of efficacy, offering a standardized and well-documented procedure through the provided script.

A key point to consider in the data collection process is our effort to source the data as directly as possible, ensuring that only the variables intentionally excluded are omitted. This approach necessitated engagement with service providers, who, at the time of data extraction, were cooperative and granted access for academic purposes without charge.

One potential drawback of this approach is the need for precision when tokenizing the data, particularly regarding the inclusion and exclusion criteria for the results. Researchers must also account for any contextual variations across the dataset.

When examining the price non-synchronization phenomenon, as proposed by Morck et al. (2000), which is suggested to stem from the R^2 of CAPM, the method outlined here offers a novel avenue for explaining and investigating price non-synchronization using non-transaction-based data. This provides an opportunity to uncover the underlying themes and motivations behind transactions before they occur. Future research could build on this analysis to explore market dynamics and investor behaviour more deeply.

In this study, the extracted data were utilised to explain asset price movements in instances where industry or market indices were insufficient. Going forward, this approach could also be adapted to explore individual preferences and market trends based on information disseminated by individuals on social media platforms.

References

Acharya, V., & Pedersen, L. H. (2005). Liquidity risk and asset pricing. *Journal of Financial Economics, 77*(2), 375-410. https://doi.org/10.1016/j.jfineco.2004.06.007

Amihud, Y., & Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics, 17*(2), 223-249. https://doi.org/10.1016/0304-405X(86)90065-6

Bhattacharya, R. (2015). Capital asset pricing model. In M. Augier & D. J. Teece (Eds.), *Palgrave Encyclopedia of Strategic Management*. Palgrave Macmillan

Black, F., Jensen, M. C., & Scholes, M. (1972). The capital asset pricing model: Some empirical tests. In M. C. Jensen (Ed.), *Studies in the Theory of Capital Markets* (pp. 79-121). Praeger Publishers

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 57-82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Chordia, T., Roll, R., & Subrahmanyam, A. (2008). Liquidity and market efficiency. *Journal of Financial Economics, 87*(2), 249-268. https://doi.org/10.1016/j.jfineco.2007.03.005

Dempsey, M. (2013). The capital asset pricing model (CAPM): The history of a failed revolutionary idea in finance? *Abacus, 49*, 62-68. https://dx.doi.org/10.2139/ssrn.2260197

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3-56. https://doi.org/10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, 18(3), 25-46.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

Grace, J. B. (2021). Instrumental variable methods in structural equation models. *Methods in Ecology and Evolution, 12*(7), 1148-1157. https://doi.org/10.1111/2041-210X.13600

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods, 41*, 924-936. https://doi.org/10.3758/BRM.41.3.924

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies, 28*(3), 650-705. https://doi.org/10.1093/rfs/hhu068

Lahouel, B.B., Zaied, Y.B., Lotfi, T., & Awijen, H. (2021). Corporate green investment and financial performance: New evidence from an instrumental variable panel quantile regression approach. *Question(s) de Management*, *6(*36), 129-139. https://doi.org/10.3917/qdm.216.0129

Lal, A., Lokhart, M., Xu, Y., & Zu, Z. (2024). How much should we trust instrumental variable estimates in political science? Practical advice based on 67 replicated studies. *Political Analysis, 32*(4), 521-540. https://doi.org/10.1017/pan.2024.2

Lee, S.C., & Eid Junior, W. (2018). Portfolio construction and risk management: Theory versus practice. *RAUSP Management Journal*, *53*(3), 345-365. https://doi.org/10.1108/RAUSP-04-2018-009

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics, 47*(1), 13-37. https://doi.org/10.2307/1924119

Long, H., Zaremba, A., & Jiang, Y. (2020). Price nonsynchronicity, idiosyncratic risk, and expected stock returns in China. *Economic Research-Ekonomska Istraživanja*, *33*(1), 160-181. https://doi.org/10.1080/1331677X.2019.1710229

Maydeu-Olivares, A., Shi, D., & Fairchild, A. J. (2020). Estimating causal effects in linear regression models with observational data: The instrumental variables regression model. *Psychological Methods*, *25*(2), 243-258. https://doi.org/10.1037/met0000226

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica*, *41*(5), 867-887. https://doi.org/10.2307/1913811

Morck, R., Yeung, B., & Yu, W. (2000). The information content of stock markets: Why do emerging markets have synchronous stock price movements? *Journal of Financial Economics*, 58(1-2), 215-260. https://doi.org/10.1016/S0304-405X(00)00071-4

Pastor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy, 111*(3), 642-685. https://doi.org/10.1086/374184

Peng, Z., & Li, X. (2018). Application of a multi-factor linear regression model for stock portfolio optimization. *Proceedings of the 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, Hunan, China, 367-370. http://dx.doi.org/10.1109/ICVRIS.2018.00096

Perold, A. F. (2004). The capital asset pricing model. *Journal of Economic Perspectives, 18*(3), 3-24. https://doi.org/10.1257/0895330042162340

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis*. Wiley. https://doi.org/10.1002/9781118391686

Roll, R. (1977). A critique of the asset pricing theory's tests: Part I: On past and potential testability of the theory. *Journal of Financial Economics*, 4(2), 129-176. https://doi.org/10.1016/0304-405X(77)90009-5

Saarinen, S. (2023). Code for the social media analysis.

Saarinen, S., Litvinenko, A. (in press). Assessing social media's influence on asset price synchronization: An econometric analysis of private information impact. *Economics and Culture.*

Shafer, G. & Vovk, V. (2019). *Game-theoretic foundations for probability and finance*. Wiley

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance, 19*(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Smith, T. A., Caligiuri, A., & Montana, J. R. (2018). Using a multiple linear regression model to calculate stock market volatility. *International Journal of Modern Trends in Technology and Science*, *57*(4), 220-224. https://doi.org/10.14445/22315373/IJMTT-V57P531