

Investor Sentiment and Stock Market Volatility: A Study Using Social Media Data and Financial Indicators

Dr. Rupinder Kaur Bhatti¹

¹Assistant professor, M.L. Memorial technical college, Moga, Punjab.

Abstract

The study analyzes the correlation between the sentiment and social-media attention toward the stock and the stock price movement by examining daily data from Twitter and the stock market for 25 stocks selected for one year starting from September 2021 to September 2022. The study accounts for the direction as well as the size of the following day's price change. Sentiment scores, proportions, and probabilities measure sentiment; tweet volume is the main indicator of attention. Descriptive statistics, correlation analysis, lagged analysis, attention-group comparisons, robust regression, stock fixed-effects models, and machine-learning methods with a strict chronological train-test procedure are used for the analysis. The results show that sentiment from social media is only marginally and inconclusively associated with later stock returns. The predictive power of classification models was found to be under the 55.12% majority class baseline and XGBoost did not perform well. By contrast, social-media attention showed a weak and positive correlation with the size of future price changes. The magnitude of absolute next-day returns rose from 2.05% in the low-attention observations to 2.38% in the high-attention observations. The relationship between tweet volume and absolute returns was positive and the relationship was robust to controlling for market variables and stock fixed effects. The results indicate that the attention paid by social media might be more indicative of the magnitude of stock-price movements than of their direction. But these findings suggest association, not causation.

Keywords: Social-media sentiment, Twitter, investor attention, tweet volume, stock returns, stock-price movement, machine learning, financial markets. markets, Twitter.

Chapter 1: Introduction

1.1 Background of the Study

The perception, interpretation, and response to information is as important to financial markets as the fundamental information. Traditional theories of finance have been ones that focus on the importance of the public information and the efficient use of information in determining asset prices. According to the Efficient Market Hypothesis, the prices of securities reflect information available to them, which means that there is minimal opportunity for abnormal returns from publicly available information (Fama, 1970). But later behavioural -finance studies have shown that investors' attention, sentiment, and psychology can affect their investment decision and the market outcome.

Investor sentiment can be used to mean the overall expectations and mood of the market participants about the financial instruments and market situation. Baker and Wurgler (2006, 2007) showed that investor sentiment can be correlated with stock return spreads and can affect the cross-section of stock prices. In the same vein, Tetlock (2007) demonstrated that one can extract information from the style of the media message when it comes to stock-market activity. The results provide some evidence that although financial market information is not confined to traditional financial statements and economic signals, textual and behavioural information can also be useful for understanding the dynamics of financial markets.

As social media has grown, there is more and faster financial opinions and information to develop and share. The social media like Twitter have opened a way for investors, analysts, companies, journalists, and other market

participants to interact with each other on a continuous basis. Social media has thus become a significant avenue for the publicly observable investor communication. Initial studies have shown that the content of financial-related Internet discussions can be linked to the stock market (Antweiler & Frank, 2004). As Twitter grew, some researchers were able to look at the sentiment and information in tweets to see if they could correlate these with the returns and movement of stocks.

According to one telling study by Bollen et al. (2011), sets of aggregated Twitter mood indicators might be correlated with future movements in the stock market. Sprenger et al. (2014) also explored stock related microblogs and discovered that microblogging on Twitter was found to carry information on the financial markets. Social-media sentiment and stock returns were examined by Sul et al. (2017), and they found evidence of a relationship between certain dimensions of the Twitter sentiment and the subsequent stock returns. Such research helped to build an understanding that data found on social media might be useful to analyse financial markets.

The link between social-media sentiment and stock returns, however, does not need to be consistent across markets, securities, time periods, or measurement approaches. For instance, Behrendt and Schmidt (2018) studied the relation of intraday sentiment and Twitter activity with individual stock-return volatility, finding that sentiment and volume of social media activity should be distinguished from one another. In a similar vein, Audrino et al. (2020) explored sentiment and attention measures as they relate to stock-market volatility and observed that attention measures could be helpful to understand realized volatility. Indeed, more recent studies have highlighted that social media attention, and sentiment can be a different source of information and are not necessarily interchangeable measures of investor behaviour (Cookson et al., 2024).

Investor attention is especially relevant since the level of attention that investors give to a security can also have information that is distinct from the sentiment of conversations around that security. Barber and Odean (2008) provided evidence that attention-grabbing events can influence investor trading behaviour and Da et al. (2011) constructed a market-based measure of investor attention and found that attention can be detected by observable information-search behaviour. Similarly, in the social media world, you can use the volume of tweets as an observable signal of the amount of public attention a company receives. Rakowski et al., (2021) explored, how investor attention to Twitter activities can be analyzed, and how the activity of the social media is relevant for information diffusion in financial markets. Also, Wu (2019) explored the relationship between social-media attention and stock returns.

Yet there is still a crucial distinction that has been overlooked—that sentiment expressed on social media may not necessarily predict the direction of movement but that the level of social-media focus may be more correlated with the magnitude of subsequent stock-price swings. A stock might see an unusually high volume of attention, due to earnings reports, surprise company news, market volatility, or other circumstances. This attention does not necessarily mean that investors are solely positive or negative. Rather, a larger focus could be associated with more disagreement, uncertainty, information arrivals, or trading volume. Thus, the focus on attention can be more telling of how much a stock will move rather than whether it will move up or down.

This is important because much existing research has focused on the predictive association between sentiment and stock returns. At the same time, the literature on attention suggests that investor attention can be related to market activity and volatility. Andrei and Hasler (2015) theoretically and empirically linked investor attention with stock-market volatility, and Audrino et al. (2020) demonstrated that sentiment and attention metrics can make different contributions to the modelling of market volatility. Thus, there is a clear difference between the return direction and the return magnitude, which can serve as an important starting point for further empirical investigations.

Now, with the advent of ever more powerful computational tools, researchers have begun studying the relationships in financial markets using machine-learning approaches. Random Forest is an ensemble method that can be used to model nonlinear relationships (Breiman, 2001), while XGBoost is an efficient gradient-boosting framework for predictive modelling (Chen & Guestrin, 2016). About asset pricing, Gu et al. (2020) showed the

potential of machine-learning methods in empirical asset pricing. This is especially the case if the financial results could be influenced by nonlinear relationships between market variables, trades, sentiment, and other information sources.

Given this context, this study examines how social-media sentiment, as captured in Twitter, relates to investor attention and subsequent stock-market performance. The study takes 25 selected stocks and observes them over a period of 1 year and creates sentiment, tweet-volume, market, return, volatility, price-range, and volume related variables. The analysis tests for the association between social-media sentiment and future stock returns, and between social-media attention (defined primarily as the number of tweets) and the size of the future stock-price movements. Moreover, machine-learning models are used to test if market and/or social-media features can enhance the prediction of next day stock-price direction.

The study uses a chronological modelling framework, to minimise the chance of information leakage between the training and testing observations. Besides the correlation-based analysis, the study is carried out using the regression models with robust inference, stock level fixed effects, and clustered standard errors. Several prediction models are also contrasted to a benchmark of the majority-class. This mix of statistics and machine learning permits the study to determine if the data available in social media is correlated with stock-market outcomes, and if its information can even be practically used.

Importantly, the study does not require that social-media sentiment be able to accurately foretell the direction of stock prices. Rather, it empirically investigates if sentiment and attention have distinct roles with respect to financial-market outcomes. This distinction offers the key driving force behind the current study and adds to the subtleties about information on social media in financial markets.

1.2 Problem Statement

With the emergence of social media, there has been a vast increase in publicly available information about companies and financial markets. Social media, especially Twitter, allows investors and other market players to voice their opinions and share information in real time. This has led to a significant research effort to try to understand – or predict – the behaviour of the stock market by using the information found on social media.

Prior research has found varying results on the association of social-media sentiment and stock returns. While some studies have theorized that Twitter sentiment carries information that can predict future market movements (Bollen et al., 2011; Sul et al., 2017), others have pointed to the significance of Twitter activity and investor attention on explaining investor volatility and trading behaviour (Behrendt & Schmidt, 2018; Audrino et al., 2020; Rakowski et al., 2021). But sentiment and attention should not be assumed to be the same phenomenon, given these differences.

One of the other issues is that research studies looking at just the direction of the return might miss considering one other potentially important outcome: stock-price change. High activity on social media can be correlated with greater attention, uncertainty and disagreement in the public sphere and/or greater information arrival without necessarily contributing to a positive or negative trend. Thus, the weak correlation between sentiment and return direction does not rule out that there is no correlation between social-media information and financial-market behaviour.

The present study tackles this issue by separately analyzing the effect of social-media sentiment and social-media attention. It examines the relation between sentiment and next stock returns and next direction of stock returns, as well as between the volume of tweets and absolute next-day returns, controlling for appropriate market variables. Machine-learning models are also employed to see if social-media and market characteristics can offer valuable predictive signals about the direction of the next day's stock price action.

1.3 Research Gap

While past studies have well covered the social-media sentiment and financial markets, there are still a few research gaps. First, many studies have been conducted on the predictive power of social media sentiment towards stock returns or return directions (Bollen et al., 2011; Sul et al., 2017). But the evidence on the consistency and the magnitude of this relationship is mixed.

Secondly, there is the conceptual difference between investor attention and investor sentiment, which are sometimes used together. Attention measures how much investors or the public pay attention to a security, while sentiment measures the tone of the information surrounding the security (whether positive, negative, or neutral). The literature on attention suggests that attention may have significant implications for trading and volatility (Barber & Odean, 2008; Da et al., 2011; Andrei & Hasler, 2015).

Third, there has been comparatively little focus on whether social-media attention is associated specifically with the size of the stock-price move, in addition to its direction. This can be supported by research into volatility (Behrendt & Schmidt, 2018; Audrino et al., 2020), but further study using a combination of social-media attention, sentiment, market controls, and predictive modelling is warranted.

Fourth, methodological variations can give false conclusions if temporal information leakage is not properly controlled. A model trained on randomly subsampled financial data can end up with information from later periods for its training sample. The present study will thus use a strict chronological train-test split for the machine-learning evaluation.

Finally, only a limited number of studies consider the traditional statistical links with machine learning prediction model and differentiate return direction and return magnitude. The current research fills this void by using correlation analysis, attention-based regression, fixed-effects modelling, robustness analysis, and machine-learning classification.

1.4 Objectives of the Study

Primary Objective

To examine the association of social-media sentiment and social-media attention, measured through Twitter activity, with the direction and magnitude of subsequent stock-price movements.

Secondary Objectives

1. To explore the relation between sentiment expressed on Twitter and returns on the stock.
2. To find out if there is a relationship between the sentiment expressed on social media and the direction of stock-price changes on the subsequent day.
3. To investigate the link between Twitter activity/tweet volume and the size of the following movement in the stock-price.
4. To test if the link between social-media attention and absolute stock returns is robust to controlling for previous returns, volatility, trading volume, and intraday price range.
5. To investigate if there is a relationship between the attention received on social media and the movements of the stock price, depending on the characteristics of the stock.
6. To investigate whether lagged social-media sentiment is associated with stock returns over different future time horizons.
7. To compare the performance of machine-learning models based on market and social-media attributes for next-day stock-price direction prediction.

8. To compare the predictive performance of machine learning models with a majority class benchmark in strict chronological train-test framework.

1.5 Research Questions

The study aims to answer the following research questions:

RQ1: Do social media sentiment on Twitter significantly influence later share returns?

RQ2: Is sentiment expressed on Twitter a valuable source of information to predict next-day stock-price movements?

RQ3: Does the level of Twitter usage predict the size of the stock price changes?

RQ4: Does the relationship between social-media attention and the magnitude of next-day stock-price movements remain significant after controlling for previous returns, volatility, trading volume, and intraday price range?

RQ5: Are there any differences in the relationship between social-media sentiment and stock returns about different return horizons?

RQ6: Do machine-learning models based on the market and social-media features perform better than a majority-class benchmark in predicting the direction of change in stock prices for the following day?

RQ7: Are there differences in the information captured by social-media attention versus sentiment that leads to differences in predicting stock-market behaviour?

1.6 Research Hypotheses

From the literature surveyed, the following hypotheses are formulated from the existing literature on investor sentiment, investor attention, and social media activity, as well as financial market volatility.

H1: Social-media sentiment is significantly associated with subsequent stock-market returns and the direction of next-day stock-price movements.

H2: Social-media attention, measured by tweet volume, is significantly positively associated with the magnitude of subsequent stock-price movements, measured by absolute next-day stock returns.

Corresponding Null Hypotheses

H0₁: Social-media sentiment is not significantly associated with subsequent stock-market returns or the direction of next-day stock-price movements.

H0₂: Social-media attention, measured by tweet volume, is not significantly positively associated with the magnitude of subsequent stock-price movements.

1.7 Significance of the Study

From a theoretical point of view as well as an empirical one, the study is important.

1.7.1 Theoretical Significance

The study has also offered new insights into the literature on behavioural finance and investor attention by separating sentiment and attention. Previous studies have shown that investor attention can affect investor trading actions and the market outcomes (Barber & Odean, 2008; Da et al., 2011). In a departure from existing research, the study considers sentiment and attention separately, establishing a rubric to check whether sentiment and attention oriented social-media activity align with different facets of the stock-market behaviour.

1.7.2 Empirical Significance

The study offers empirical evidence in a multi-stock data set with 25 companies for a period of about one year. It does not only consider information from social media for the directional prediction, but also the absolute price movement. This would allow for a more detailed analysis of the relevance of social-media activity as a financial issue.

1.7.3 Methodological Significance

The study uses both traditional statistical analysis as well as machine-learning techniques. A number of approaches to analysing the relationship are used: correlation analysis, robust regression, stock fixed effects, clustered standard errors, and machine-learning classification. The chronological train-test split also allows the use of a more realistic measure of predictive performance in a time-dependent financial context.

1.7.4 Practical Significance

The results could be of benefit to researchers, investors, financial analysts, and practitioners with an interest in alternative information sources. In particular, knowing whether the relation between social-media activity and market movement is more closely related to the intensity of the market movement than to the direction of the movement can help determine whether the use of social media is as a directional trading signal or as an indicator of increased market intensity.

1.8 Scope of the Study.

The study is based on a collection of Twitter-related data for 25 selected stocks for one year to investigate the relationship between the information in social media and stock-market behaviour. The analysis considers sentiment, volume of tweets, past stock performance, momentum, volatility, trading volume, and intraday price action.

The primary outcome measures are:

- 1) Next day stock return to look at return relationships.
2. Direction of stock-prices the next day, which is a binary classification.
3. Absolute next day return, as a measure of the size of subsequent stock-price movement.

The machine-learning component compares models with a stringent separation of the observations between training and testing time. Relationships are explored in the statistical part, both unconditional and conditional on market variables that are relevant.

The study relates to associations and predictive performance and not causal identification. Thus, one should not conclude from the results that there is any direct link between the activity on Twitter and the movement in the stock price.

1.9 Organization of the Study

The rest of the research paper is structured in the following manner.

Chapter 2 reviews the literature on investor sentiment, investor attention, social media, Twitter activity, stock returns, stock-market volatility, and machine-learning methods in financial markets.

Chapter 3 explains the research methodology that involves research design, data source, sample selection, sentiment measurement, feature engineering, dependent variables, independent variables, statistical techniques, machine learning models, and model validation procedures.

Chapter 4 provides the empirical results: descriptive statistics, sentiment-return analysis, attention-return analysis, lagged relationships, regressions, fixed-effects analysis, robustness analysis, and machine learning performance.

Chapter 5 presents the results and interprets them with reference to the existing literature and in terms of theory and practice.

Chapter 6 reflects on the key findings, conclusion, implications, limitations, and further research.

Chapter 2: Review Of Literature

2.1 Introduction

Investor behaviour and stock-market performance have received much attention in financial economics. The theory of traditional asset pricing theory, dominated by the Efficient Market Hypothesis (see Fama, 1970), indicate that the value of securities must reflect publicly available information very quickly, and that there is little opportunity to make consistent predictions based on such information. But from the behavioural finance point of view, it is seen that investor sentiment, limited attention, and psychological biases can affect investment decisions and market outcomes (Baker & Wurgler, 2006, 2007; Hirshleifer & Teoh, 2003).

With the advent of social media, there is now a wealth of investor-created information that is freely available. Twitter in particular offers investors and others a continuous stream of time-stamped communications, in which they voice opinions, respond to events, and share information about specific securities. Thus, data from social media have become a major source of study as proxies for investor sentiment, attention, information spreading, and stock market behaviour (Bukovina, 2016).

So far, the literature has focused on two aspects of social-media information—sentiment and attention. Sentiment indicates whether the communication is positive, negative, or neutral, while attention indicates whether there is interest or activity in a stock. There is a relationship between these concepts, but they correspond to different modes of behaviour. It is important to note that a stock can get a lot of attention for the good news but can also get a lot of attention for the sad news and high attention does not necessarily mean the price will move either up or down.

The difference is relevant for the present study as the stock-market behaviour can also be viewed in two ways: direction and magnitude. A social media metric could have a weak power of prediction on whether a stock price will rise or fall but could be still linked to the magnitude of the resulting price change. Thus, the literature on sentiment, social-media activity, investor attention, volatility, and machine-learning-based financial prediction is explored.

2.2 Investor Sentiment and Stock-Market Behaviour

Investor sentiment is one of the key elements in behavioural finance. Baker and Wurgler (2006) showed that investor sentiment can affect the cross-section of stock returns and Baker and Wurgler (2007) stressed the more general role of sentiment in understanding stock-market behaviour. These studies indicate that the expectations and psychological attitude of investors can influence asset prices, especially in a situation of uncertainty in the information environment.

Another way of measuring investor sentiment is using textual information. Tetlock (2007) studied the correlation between media negativity and stock-market results and concluded that media content does carry information pertinent to the stock market. Tetlock et al. (2008) further demonstrated that textual information can provide information about firms' fundamentals. All these studies suggest that financial communication language may include economically relevant information.

But do sentiment measures provide good short-term return prediction? Investor sentiment can be related to the reaction to information as well as to expectations for the future. As a result, there is no clear or consistent relationship between sentiment and future returns that is independent of time, security, market conditions, or

sentiment measurement. This is especially important if sentiment is based on social-media communication, in which the message could come from someone with varying degrees of information or have a large amount of "noise" in it.

2.3 social media as a source of financial information

Social media has established an added information environment where market participants can communicate continuously. Social media differs from traditional financial media in that it has provided individual investors the opportunity to create and share information directly. Because of its size and frequency, social-media data can be analyzed for insights into investor behaviour and market dynamics.

Bukovina (2016) studied the application of big data in capital markets, specifically focusing on those relating to investor sentiment, attention, and financial-market analysis. Previous studies of Internet-based financial communication also suggest that the Internet can include market relevant information in the discussion. For instance, Antweiler and Frank (2004) studied the Internet stock message boards and discovered that there was information associated with market activity within the Internet discussions.

The literature was later expanded to include research done on Twitter. Sprenger et al. (2014) examined stock microblogs and revealed that Twitter communication is useful for the financial-market activity. Their results validate the use of social-media data as an alternative source of information to examine investor behaviour.

However, by no means are information content and predictive ability the same thing. A social-media variable can be associated with current market conditions but may not be a good predictor of future returns. This separation is an encouragement to consider lagged social-media metrics in the current study.

2.4 Twitter Sentiment and Stock Returns

A study that made a significant contribution in this field is Bollen et al. (2011), in which the authors look at the possibility of using Twitter-based mood measures to predict the movement of the stock-market. Their results indicated that some aspects of overall Twitter mood were predictive of future market activity and stimulated a growing interest in using social media to predict financial outcomes.

Later the studies more nuancedly showed. Sprenger et al. (2014) reported the information content of stock-related microblogs and Sul et al. (2017) analysed the data on social-media sentiment as a predictor of stock returns. Such research takes the form of evidence which suggests that social-media sentiment may include information pertinent to financial markets.

The evidence is not consistently favourable for good directional prediction, however. Behrendt and Schmidt (2018) investigated Twitter activity and investor sentiment with respect to the individual level stock-return volatility and found the need to account for the market outcome of interest. Their results confirm that the direct connection between Twitter activity and stock-price movements is not to be taken for granted.

But the literature is mixed about whether sentiment on social media is "predictive" of stock prices. Rather, specific characteristics of social media must be identified, as well as specific outcomes of financial markets must be identified.

The present study aims at tackling this question by focusing on sentiment's association with subsequent returns and the direction of the next day's stock-price change.

2.5 Investor Attention and Financial Markets

Investor attention is another aspect of the behaviour of the market that is distinct from sentiment. Sentiment indicates where investor communication lies while attention shows how much investors are concerned with a specific security or event.

The theoretical foundation of investor attention can be associated with limited information processing. Hirshleifer and Teoh (2003) pointed out that investors' information processing abilities are constrained, suggesting that heterogeneity in attention may have an impact on the use of information in making financial decisions.

There is empirical evidence to support the importance of attention in financial markets. Barber and Odean (2008) showed that alluring events can affect investors' trading behaviour. Da et al. (2011) constructed a measure of investor attention that is observable and demonstrated that attention matters in stock-market behaviour.

The results have significant consequences on social-media analysis. When there is a significant jump in the number of tweets, it could be because of a significant corporate event, unexpected information, uncertainties in the market, controversy, or the increased interest by investors. Thus, attention can be linked to an increase in market activity, but it does not imply anything about the sign of the following return.

2.6 Social-Media Attention and Stock-Market Volatility

The connection between investor attention and market volatility offers an essential theoretical basis to differentiate between price direction and price-movement magnitude.

Investors' attention and market volatility were studied by Andrei, Hasler (2015) and they discovered evidence that attention is related to future volatility. They suggest that more investor attention can be associated with greater financial market volatility.

Audrino et al. (2020) also analysed sentiment and attention metrics as they relate to stock-market volatility. Significantly, their study shows that sentiment and attention can be differently related to financial-market outcomes.

The attention and information diffusion have also been studied in the context of social media. Rakowski et al. (2021) examined Twitter activity, investor attention and the spread of information and found that social-media activity is relevant to investor attention. In a similar study, Wu (2019) studied the social media attention and stock return.

Cookson et al. (2024) provide further evidence highlighting the importance of social signals in financial markets.

These studies all indicate that social-media attention could be useful as an informational clue as to the strength of market activity, even though it does not offer a consistent directional indicator. This gives rise to our focus in the current research on the absolute next-day stock returns and tweet volume.

2.7 Media Attention and Information Diffusion

Attention is crucial, but not just on social media. Fang and Peress (2009) studied the media coverage and stock returns and demonstrated that the differential in media attention can have implications for the financial-market outcomes.

The literature points to the fact that more exposure to information can impact the participation of investors and processing of information. In a social-media landscape, more discussion will further speed the spread of information since investors are able to quickly respond to and share information.

Rakowski et al. (2021) offer evidence that the use of Twitter is related to investor attention and the spread of information. This indicates that social-media activity could be a good measure of how much attention is being paid to a security.

But the focus on a greater number can come from positive and negative events. Thus, such attention does not automatically indicate the direction in which price will move. This gives some justification for looking at attention in the context of absolute returns, as well as signed returns.

2.8 Sentiment, Attention and Volatility – Integrated View

From the existing literature, there is a clear indication to separate sentiment from attention conceptually.

Sentiment is the answer to the question:

What are investors saying with their investments?

Attention answers:

What is their level of investor interest in the stock/Information?

These dimensions can be concurrent, but do not necessarily have to result in the same market outcome. For instance, a stock can have extremely high Twitter activity, whether for good or sad news. A shared trait is the focus, not necessarily the slant, of the attention.

The difference is the reason for why sentiment on social media strong directional predictive ability might not have, but attention on social media might be related to bigger subsequent price changes.

This is in line with the literature on investor attention and volatility (Andrei & Hasler, 2015; Audrino et al., 2020). For the purposes of the present study, sentiment and attention are hence viewed as explanatory dimensions than as a compounded indicator of social media.

2.9 Machine Learning and Financial Prediction

Machine-learning methods have become increasingly relevant in empirical asset-pricing research as they can uncover nonlinear interactions and relationships between large numbers of variables. Gu et al. (2020) showed how machine learning can be employed in empirical asset pricing and highlighted the need to test predictive ability using proper out-of-sample methods.

Random Forest is an ensemble learning technique that can model nonlinear relationships using several decision trees (Breiman, 2001). XGBoost is an extension of tree-based learning using gradient boosting and has been applied to a wide variety of structured prediction problems (Chen & Guestrin, 2016).

When applied to financial research based on social media, machine learning can be used to perform a test of whether sentiment and attention can offer incremental predictive information on top of market variables.

Predictive performance should not be misinterpreted for statistical association, however. It is possible for a variable to be statistically associated with an outcome and not to provide useful out-of-sample predictions. For this reason, the present study is not a comparative approach that aims to evaluate the performance of machine-learning models as distinct theoretical hypotheses.

2.10 Volatility and Measurement of Stock-Price Movement

The returns of the financial markets have been known to have varying variance and extreme observations. The ARCH model was devised by Engle (1982) to capture conditional heteroskedasticity which was further developed by Bollerslev (1986) under the GARCH framework. The distribution of realized stock-return volatility was also studied by Andersen et al. (2001).

These studies show how important it is to pay attention to the difference between the magnitude of financial market moves and the direction of those moves.

Hence, the present study takes the next day absolute return as a proxy for the magnitude of the price change:

$$|R_{t+1}|$$

This is a measure that rewards positive and negative movement to the same extent. A plus and minus 5% return are both 5% price changes, for instance.

This is especially relevant when looking to see if social-media attention correlates to greater price activity in the future, whether positive or negative.

2.11 Methodological Considerations

Heteroskedasticity, extreme observations, and dependence among observations from the same security are common characteristics of financial and social-media data sets. To accomplish better statistical inference in the presence of heteroskedasticity, White (1980) introduced heteroskedasticity-consistent covariance estimation.

Stock-level dependence must also be considered in the present study because there are repeated observations for individual stocks. When observations in groups may be correlated, see Cameron and Miller (2015) for guidance on cluster-robust inference.

There is also a multiple-testing issue with the use of multiple stocks and several statistical tests. An innovative approach to correcting for multiple comparisons called false-discovery-rate control was developed by Benjamini and Hochberg (1995). Similarly, Harvey and Liu (2016) emphasized the need to consider false discoveries in empirical research in Finance.

These methodological issues provide the basis to advocate the use of strong inference and careful interpretation of individual stock-level results in the current study.

2.12 Research Gap

As noted in the literature surveyed above, there are a number of areas which have not been sufficiently covered.

First, prior research has focused on the link between social-media sentiment and stock returns, but the evidence on a directionally consistent prediction on the next day is mixed.

Second, sentiment and attention are often talked of together but are different attributes of investor behaviour. Sentiment records the direction of communication; attention records the strength of communication.

Third, there is a relative lack of focus on return prediction as opposed to potential link between social-media activity and the size of future price changes.

Fourthly: Statistical association and predictive performance are various kinds of evidence. So, comprehensive evaluation should involve not only statistical inference, but also out-of-sample prediction.

The present study aims to fill these gaps by analysing 25 stocks while separately analysing the volume of tweets and the sentiment of social media. Sentiment is measured with respect to performance in the following days and next day price changes; tweets volume is observed with respect to absolute next day returns. The analysis is based on correlation analysis, lagged relationships, robust regression, stock fixed effects, and machine-learning models.

Thus, the study is focused on changing the research question from the merely "can social media 'predict stock prices'?" to a more specific question:

> Do sentiment changes on social media have predictive power about future stock-price changes, and does social-media coverage have predictive power about the magnitude of those changes?

This difference is at the very heart of the research gap which the study captures.

Chapter 3: Research Methodology

3.1 Research Design

In the present study, we use a quantitative empirical research design to investigate the impact of social-media sentiment and social-media attention on future stock-market activities. The study uses descriptive statistics, correlation analysis, regression analysis, fixed-effects estimation, and machine-learning techniques to assess the statistical association and the out-of-sample predictive performance.

The methodology is based on two aspects of the stock-market behaviour. The first study considers if social-media sentiment is related to future stock returns and direction of the next day's price change. Second, it examines whether the level of social-media attention (in terms of number of tweets) is related to the size of the ensuing stock-price movements (in absolute terms of next-day returns).

This separation allows this research to investigate whether social-media data is more useful for identifying the direction of price moves or for identifying periods of increased market-movement intensity.

3.2 Data and Sample

The study uses daily stock-market and Twitter/social-media data for 25 selected stocks over the study period from 30 September 2021 to 29 September 2022.

There are 6,275 stock-day observations, 252 observations for each stock. The data set was checked for stock-date duplications and chronological inconsistencies. There were no duplicated stock-date observations or stock-level date-order issues found.

The number of observations the classification dataset contains is 6,275 after building the next-day direction target. The classification period is 30 September 2021 – 28 September 2022.

3.3 Data of Social Media and Sentiment Measurement

Information from social media is integrated using day-to-day measures of sentiment and Twitter activity in relation to specific stocks. The volume and sentiment attributes of the selected stock's tweets are considered in the study.

Sentiment is measured through various indicators such as:

1. average sentiment.
2. net sentiment.
3. proportional simplexes and sizes: positive, negative, and neutral.
4. probabilities of positive, negative, and neutral sentiment; and
5. sentiment intensity measures.

Sentiment measures are compiled on a daily stock level. To reduce the possibility of using contemporaneous information when explaining future market outcomes, the principal predictive models use lagged social-media variables.

The main lagged social-media variables are:

$$\text{Sentiment}_{t-1}$$

and

$$\text{TweetCount}_{t-1}$$

where $t - 1$ represents the previous trading day.

3.4 Measuring Social-Media Attention

We measure social-media attention by the number of relevant tweets of a stock in a day, denoted by tweet volume.

The number of tweets is utilized as an attention measure, as the more tweets that are sent, more people are discussing or interested in a security. The natural logarithm of the attention measure is also entered into the regression analysis:

$$\text{LogTweetCount}_{t-1} = \ln(1 + \text{TweetCount}_{t-1})$$

The logarithmic transformation will make highly active observations have less impact, and the analysis will be able to account for the highly skewed distribution of tweet volume.

There are 4,167 observations in the tweet-count data that have sentiment information. There are about 15.23 tweets on average with a median of 5 and a maximum of 462 tweets, which suggests there is an extremely elevated level of variability of social media activity across stock-days.

3.5 Stock-Market Variables

The dependent and control variables are based on daily information from the stock-market.

3.5.1 Next-Day Return

The key measure of return is the stock price difference on the next day on which it trades.

$$R_{i,t+1} = \frac{P_{i,t+1} - P_{i,t}}{P_{i,t}} \times 100$$

where $P_{i,t}$ represents the closing price of stock i on day t .

The next-day return is tested to see if there is any relationship between social-media sentiment and the stocks' returns.

3.5.2 Next-Day Price Direction

For the classification analysis, the next-day return is converted into a binary directional variable:

$$\text{Direction}_{i,t+1} = \begin{cases} 1, & \text{if } R_{i,t+1} > 0 \\ 0, & \text{if } R_{i,t+1} \leq 0 \end{cases}$$

3.5.3 Absolute Next-Day Return

To measure the magnitude of subsequent stock-price movements, the study uses absolute next-day return:

$$\text{AbsReturn}_{i,t+1} = |R_{i,t+1}|$$

This is to eliminate the sign of the return but keep the magnitude. Therefore, a positive and negative return of the same magnitude will be considered as equal in size.

This is the main dependent variable to analyse the relationship between social-media attention and the intensity of the stock-price movement.

3.6 Control Variables

The regression analysis includes a set of market variables to diminish the likelihood that the relationship between attention in social media and future price changes is merely what was occurring in the market anyway.

The main controls are:

1. Lagged daily return
2. Rolling 20-day volatility

3. Lagged trading volume

4. Intraday price range that is lagged

These variables reflect recent price action, recent volatility, trading volume, and intraday price action.

The controlled specification is as follows:

$$|R_{i,t+1}| = \beta_0 + \beta_1 \text{LogTweetCount}_{i,t-1} + \beta_2 \text{Return}_{i,t-1} + \beta_3 \text{Volatility}_{i,t-1} + \beta_4 \text{Volume}_{i,t-1} + \beta_5 \text{IntradayRange}_{i,t-1}$$

The coefficient β_1 is the principal coefficient of interest.

3.7 Feature Engineering

Feature engineering was done considering the characteristics of social media and recent market conditions.

The final classification data set includes 33 explanatory variables, such as lagged sentiment variables, tweet volume variables, sentiment ratios, sentiment probabilities, interaction variables, historical returns, momentum, volatility, price-range measures, measures of trading volume, and tweet-availability information.

Market related features include:

- lagged returns.
- 3-, 5-, and 10-day momentum.
- 5-, 10-, and 20-day volatility.
- intraday range.
- high-low range.
- open-close range.
- closing-price position.
- lagged trading volume.
- volume change; and
- volume ratios.

Social-media variables include sentiment measures with a time lag, sentiment probabilities, tweet volume, and sentiment related interaction variables.

Lagging the explanatory variables is done so that the predictive models have information to use before the future return that is being evaluated.

3.8 Descriptive and Correlation Analysis

Descriptive statistics are first employed to analyze sentiment, the number of tweets, returns, and absolute returns.

Pearson correlation is used to assess the linear association, and Spearman correlation is used as a rank-based measure which is less sensitive to a strictly linear relationship. The sentiment-return relation is separately discussed for each of the 25 stocks. This enables the study to determine if any relationship found is pervasive across securities or a handful of stocks. Sentiment at various intensities is also analysed and subsequent performance is compared between sentiment groups.

3.9 Lagged Relationship Analysis

The study tests various future time horizons to determine if sentiment provided by social media is related to future, not just contemporaneous returns.

The analysis considers:

- 1-day future return.
- 2-day future return.
- 3-day forward return; and
- 5-day future return.

Pearson and Spearman correlation are computed for sentiment and the respective future return for each horizon. This analysis offers an extra check on the robustness of sentiment as a predictor of future stock returns.

3.10 Attention Group Analysis

Tweet volume is also segmented into attention groups – that is, whether the subsequent price changes are different for stocks that receive greater or lesser amounts of social-media attention.

The attention analysis looks at low, medium, and high attention observations, in terms of both tweet volume and the market outcomes that followed.

Besides signed returns, absolute next-day returns are analyzed. Descriptive statistics and the Welch two sample comparison for high and low attention observations are used to examine for group-level differences.

The use of absolute returns enables the study to identify if high attention observations have greater changes, both positive and negative.

3.11 Regression Analysis

The relationship between Social Media attention and the size of subsequent stock-price moves is investigated using Ordinary Least Squares (OLS) regression.

Financial data may be heteroscedastic, so the White (1980) robust-inference standard errors, consistent with the presence of heteroskedasticity, are used.

The main regression model measures the relationship between the log volume of tweets and the absolute next-day return holding constant the previous day's return, past volatility, trading volume, and intraday range.

An equivalent to the regression is also analysed, which has been standardized to make the variables more comparable across different scales.

3.12 Stock Fixed-Effects Model

A stock fixed-effects specification is estimated as a robustness check.

The fixed-effects model includes unobserved, time-invariant attributes that vary across the 25 stocks. The model can be represented as:

$$|R_{i,t+1}| = \alpha_i + \beta_1 \text{LogTweetCount}_{i,t-1} + \beta X_{i,t-1}, t, t$$

where α_i are stock specific fixed effects and X is a set of control variables.

The standard errors are grouped at the stock level. This specification examines the relationship between social-media attention and the size of the price move after controlling for the persistent difference across securities.

3.13 Machine-Learning Models

Machine-learning approaches are used to assess if social-media information contains out-of-sample forecasting power of stock-price movement the following day.

The classification models are:

- Logistic Regression.

- Decision Tree.
- Random Forest; and
- Gradient Boosting.

These models are tested using market variables as well as social-media variables to determine whether the social-media variables provide better predictive power.

A regression analysis with XGBoost is also included to assess the prediction of the stock-market returns.

Thus, machine learning is not a replacement for statistical inference but rather as a predictive validation component.

3.14 Chronological Train-Test Split

To prevent leakage of information from future observations into the training sample, a strict chronological division is used.

The classification data set is split into two parts:

Dataset	Observations	Period
Training	5,025	30 Sep 2021 – 19 Jul 2022
Testing	1,250	20 Jul 2022 – 28 Sep 2022
Total	6,275	30 Sep 2021 – 28 Sep 2022

The training observations are earlier in time than the testing observations. Random shuffling is not used. The training and testing features are imputed and standardized following procedures suitable for the training features and the resulting transformation is applied to the test features.

3.15 Evaluation Metrics

The performances of the classifications are assessed with:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

Balanced Accuracy

Balanced accuracy is used since the proportions of the observations towards the upward and downward directions are not precisely equal.

ROC-AUC

The model's discriminative power on the probability thresholds is judged by ROC-AUC. A simple majority class baseline is also provided to see if machine-learning models give any improvement over the simple baseline.

The student will evaluate the quality of data through validation techniques.

There were several validation procedures before final analysis.

The dataset was checked for:

- duplicate stock-date observations.
- chronological ordering for stocks.
- missing values in features/targets for models.
- consistency between current returns and next day target returns; and
- stock-level observation counts.

The final validation did not result in any duplicate stock-date observations or any stocks with date-order problems. The underlying return series were also tested against the target return you can construct for next day with a relevant difference of 0.0.

These checks were made to minimize the chance for data duplication, time-shift mismatch, or errors in the construction of the targets.

3.17 Statistical Significance

Conventional significance levels are used to determine statistical significance, with special consideration of the 5% level.

Several stocks and relationships are analyzed so each important statistically significant finding is viewed with caution. The study makes a distinction between findings for individual stocks versus trends that survive being pooled with other stocks and robustness checks.

The analysis focuses on the direction, size, consistency, and robustness of estimated relationships and not simply on statistical significance.

Chapter 4: Results

4.1 Introduction

This chapter presents the empirical findings of the study examining the relationship between social-media sentiment, social-media attention, and subsequent stock-market behaviour. The analysis is based on 25 selected stocks observed over the study period. The results are presented through descriptive analysis, correlation analysis, lagged relationships, attention-based analysis, regression models, and machine-learning models.

The results are organized to distinguish between two dimensions of market behaviour: direction of the next-day return and magnitude of the next-day price movement. Attention is given to whether social-media sentiment provides useful information about return direction and whether tweet activity is associated with the intensity of subsequent market movements.

4.2 Daily Social-Media Activity

The daily volume of social-media activity varied throughout the study period. Although tweet activity remained within a moderate range on most trading days, several pronounced spikes were observed during particular periods. These spikes indicate periods of unusually high attention toward the selected stocks.

The variation in daily tweet activity provides evidence that social-media attention was not constant over time. Instead, attention increased on certain dates, potentially reflecting news events, market developments, or heightened investor interest.

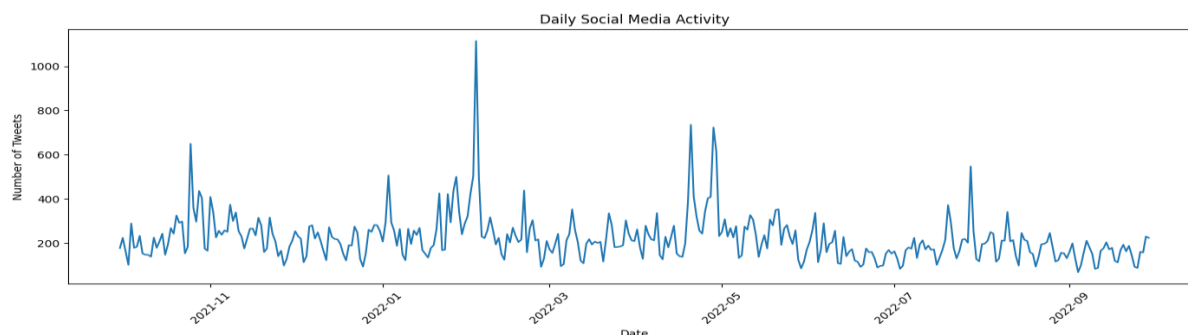


Figure 1. Daily Social Media Activity

4.3 Distribution of Next-Day Stock Returns

The distribution of next-day stock returns was concentrated around zero, while exhibiting substantial dispersion in the tails. The complete dataset contained 6,275 observations, with a mean next-day return of approximately -0.098% and a standard deviation of 3.201% .

The minimum observed next-day return was -35.12% , while the maximum was 29.55% . These values indicate the presence of extreme market movements within the sample. The median return was approximately -0.077% , suggesting that typical daily movements were small despite the presence of large positive and negative observations.

For visual clarity, the figure presents the central 99% range of the return distribution.

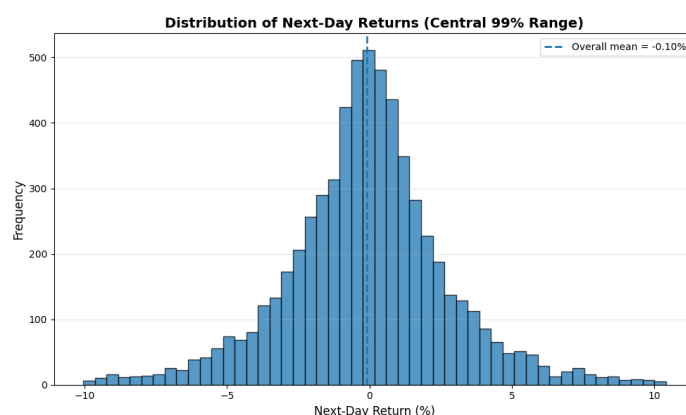


Figure 2. Distribution of Next-Day Returns (Central 99% Range)

The histogram presents the central 99% range of next-day stock returns. The dashed vertical line indicates the overall mean next-day return of approximately -0.10% .

Table 1 Descriptive Statistics of Next-Day Stock Returns

Statistic	Next-Day Return (%)
Observations	6,275
Mean	−0.0978
Standard Deviation	3.2011
Minimum	−35.1166
Q1	−1.6552
Median	−0.0767
Q3	1.3878
Maximum	29.5529

The distribution demonstrates that stock returns were highly variable and contained substantial extreme observations. This characteristic is relevant when interpreting both the regression and machine-learning results.

4.4 Tweet Data Coverage Across Stocks

Tweet availability differed across the 25 stocks. Several stocks had tweet data covering the entire trading period, whereas other stocks had lower coverage.

Stocks such as TSLA, NIO, TSM, PG, MSFT, AAPL, and AMZN had approximately 100% tweet-data coverage. In contrast, stocks such as NOC and F had approximately 9.6% coverage. This variation is important because the social-media analyses are based on days for which tweet information was available.

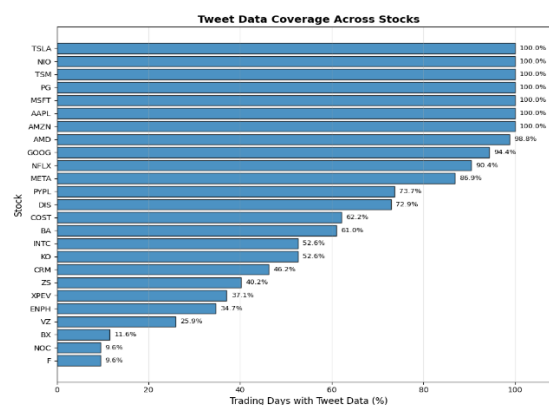


Figure 3. Tweet Data Coverage Across Stocks

The figure shows the percentage of trading days for which tweet data were available for each stock.

The unequal coverage also indicates that social-media attention was more concentrated for some securities than for others.

4.5 Average Sentiment by Stock

Average sentiment differed across the selected stocks. The highest average sentiment scores were observed for NOC and F (0.191), followed by NIO (0.156), BX (0.123), and XPEV (0.121). In contrast, several stocks showed negative average sentiment, with NFLX recording the lowest average sentiment (−0.072).

The results demonstrate that the overall tone of social-media discussions was not uniform across companies.

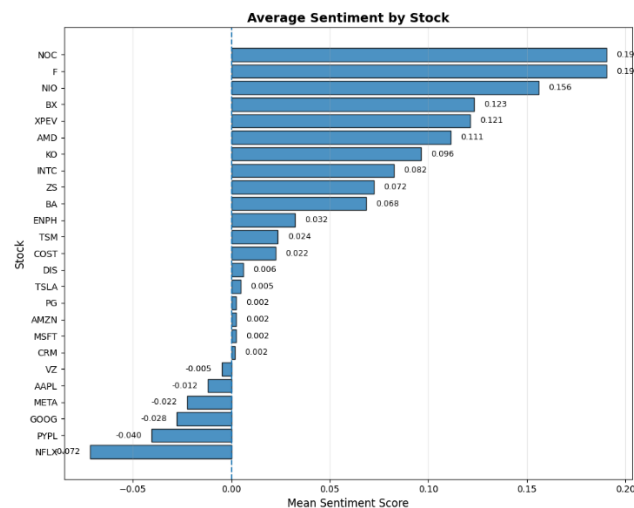


Figure 4. Average Sentiment by Stock

The figure presents the mean social-media sentiment score for each stock. Positive values indicate more positive sentiment, while negative values indicate more negative sentiment.

4.6 Average Tweet Volume by Stock

Social-media attention varied across stocks. TSLA had by far the highest average tweet volume, with approximately 119.2 tweets per trading day, followed by TSM (30.1) and AAPL (16.3). Several stocks recorded average daily tweet volumes below two tweets.

This substantial variation demonstrates that attention was highly uneven across securities and provides a useful basis for examining whether greater social-media attention is associated with subsequent market movements.

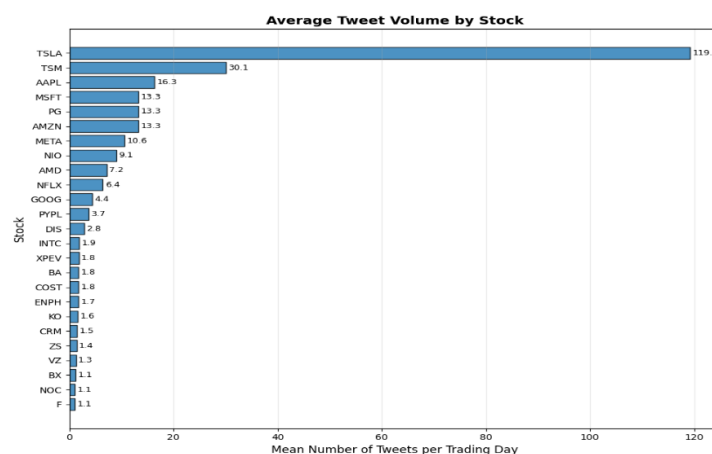


Figure 5. Average Tweet Volume by Stock

The figure presents the average number of tweets per trading day for each stock.

4.7 Social-Media Sentiment and Next-Day Returns

The overall relationship between social-media sentiment and next-day stock returns was weak. The Pearson correlation between sentiment and next-day return was approximately 0.025, indicating an almost negligible linear association.

The scatterplot also shows that observations were widely dispersed around the horizontal zero-return line. Positive and negative sentiment values were associated with both positive and negative subsequent returns rather than displaying a clear directional pattern.

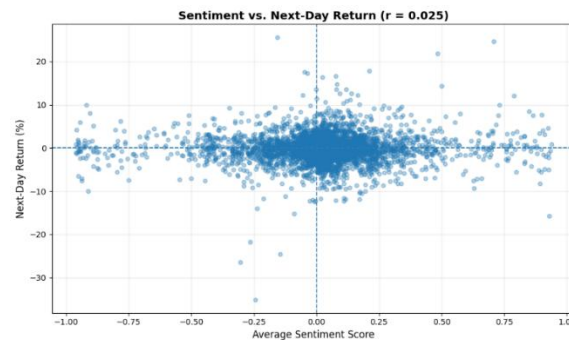


Figure 6. Sentiment vs. Next-Day Return

The scatterplot illustrates the relationship between average social-media sentiment and next-day stock returns. The reported Pearson correlation is approximately 0.025.

The overall Pearson correlation was 0.0272 ($p = 0.0796$) in the corresponding sentiment-return analysis, while the Spearman correlation was only 0.0093 ($p = 0.5486$). These results provide little evidence of a meaningful monotonic relationship between sentiment and subsequent returns.

4.8 Sentiment Intensity and Next-Day Returns

To examine whether the strength of sentiment was associated with subsequent returns, observations were divided into five sentiment-intensity groups ranging from extremely negative to incredibly positive.

The mean next-day returns did not follow a consistent increasing or decreasing pattern across the groups. The Neutral group recorded the highest mean return at 0.0482%, whereas the Negative group recorded the lowest mean return at -0.2808%. The Very Positive group recorded a mean return of -0.0229%.

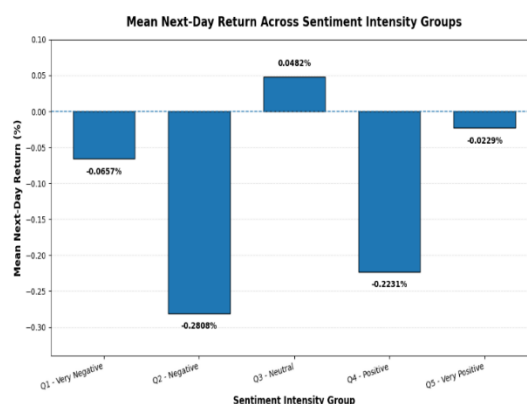


Figure 7. Mean Next-Day Return Across Sentiment Intensity Groups

Note: The figure presents mean next-day returns across five sentiment-intensity groups, ranging from extremely negative to incredibly positive sentiment.

Table 2 Sentiment Intensity and Next-Day Returns

Sentiment Group	N	Mean Sentiment	Mean Next-Day Return (%)	Up Days (%)
Q1 – Very Negative	847	−0.3463	−0.0657	49.11
Q2 – Negative	827	−0.0504	−0.2808	47.40
Q3 – Neutral	829	0.0061	0.0482	49.94
Q4 – Positive	827	0.0489	−0.2231	46.19
Q5 – Very Positive	837	0.3407	−0.0229	48.15

The absence of a systematic pattern across sentiment groups further suggests that sentiment intensity alone does not provide a reliable indicator of the direction of the following day's stock return.

4.9 Lagged Sentiment and Future Stock Returns

The study further examined whether sentiment was associated with returns over different future horizons. Pearson and Spearman correlations were calculated for one-, two-, three-, and five-day future returns.

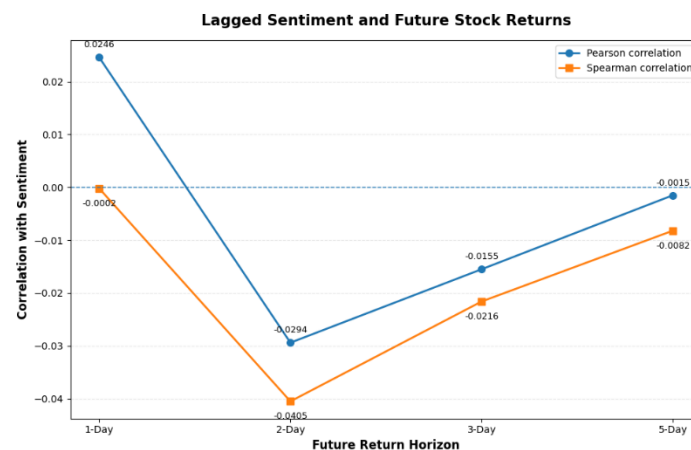


Figure 8. Lagged Sentiment and Future Stock Returns

The figure compares Pearson and Spearman correlations between lagged social-media sentiment and future stock returns across one-, two-, three-, and five-day horizons.

Table 3 Lagged Sentiment and Future Returns

Future Horizon	N	Pearson Correlation	Pearson p-value	Spearman Correlation	Spearman p-value
1-Day	4,167	0.0246	0.1121	−0.0002	0.9895

Future Horizon	N	Pearson Correlation	Pearson p-value	Spearman Correlation	Spearman p-value
2-Day	4,148	-0.0294	0.0585	-0.0405	0.0090
3-Day	4,136	-0.0155	0.3188	-0.0216	0.1641
5-Day	4,102	-0.0015	0.9244	-0.0082	0.6007

The correlations remained exceedingly small across all horizons. Although the two-day Spearman correlation was statistically significant, its magnitude was only -0.0405 , indicating a very weak association. Overall, the lagged analysis does not provide evidence of a stable or economically meaningful predictive relationship between sentiment and future stock returns.

4.10 Stock-Wise Sentiment–Next-Day Return Correlations

The relationship between sentiment and next-day returns was also examined separately for each stock. The results showed considerable variation in the sign and magnitude of correlations.

Nine stocks displayed positive Pearson correlations, while 16 displayed negative correlations. However, only NIO showed a statistically significant Pearson correlation at the 5% level, with a correlation of -0.1462 ($p = 0.0208$).

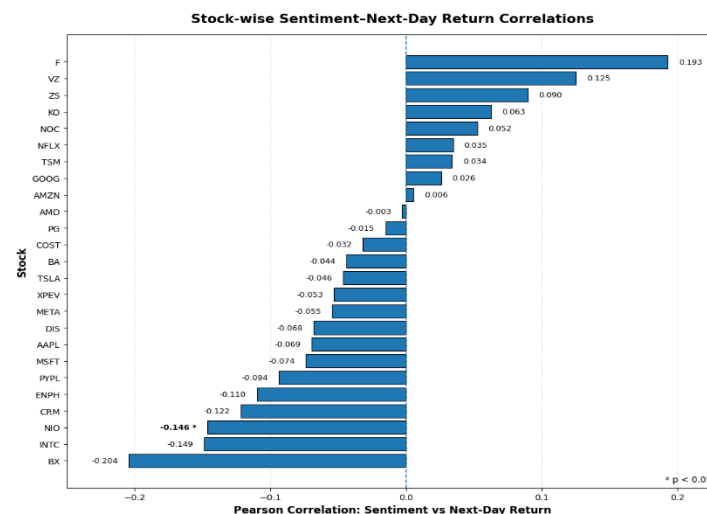


Figure 9. Stock-wise Sentiment–Next-Day Return Correlations

The figure presents Pearson correlations between social-media sentiment and next-day stock returns for each stock. The asterisk indicates statistical significance at $p < 0.05$.

Table 4 Summary of Stock-Wise Sentiment–Return Correlations

Measure	Result
Stocks with positive Pearson correlation	9
Stocks with negative Pearson correlation	16

Measure	Result
Stocks significant at $p < 0.05$	1
Significant stock	NIO
NIO Pearson correlation	-0.1462
NIO p-value	0.0208
Average directional accuracy	50.02%
Stocks with directional accuracy $> 50\%$	11

The stock-wise findings therefore reinforce the conclusion that sentiment does not demonstrate a consistent relationship with next-day return direction.

4.11 Tweet Attention and Next-Day Market Movement

The analysis next shifted from sentiment to social-media attention, measured by tweet volume. Observations were divided into low-, medium-, and high-attention groups. A positive descriptive pattern was observed when the dependent variable was the absolute next-day return, which represents the magnitude of the price movement regardless of whether the movement was positive or negative.

The mean absolute next-day return increased from 2.05% in the low-attention group to 2.21% in the medium-attention group and 2.38% in the high-attention group. This pattern indicates that higher levels of social-media attention are associated with larger subsequent stock-price movements, regardless of whether the movement is positive or negative.

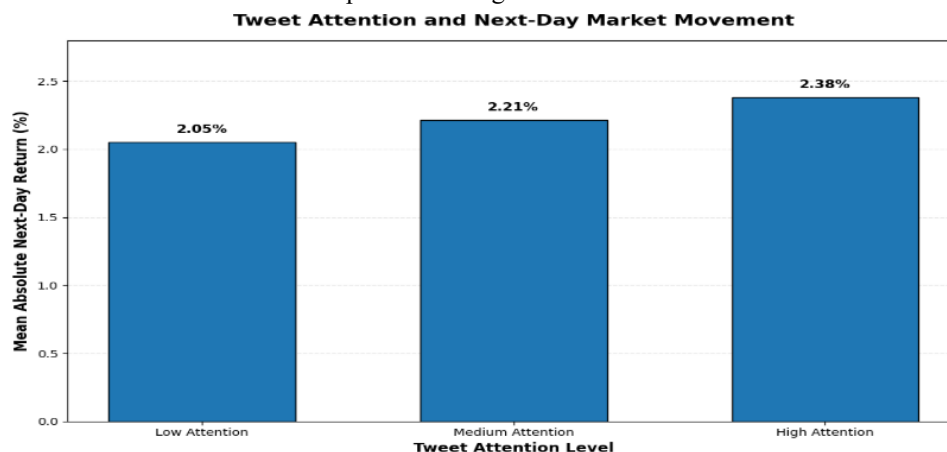


Figure 10. Tweet Attention and Next-Day Market Movement

The figure compares mean absolute next-day stock returns across low-, medium-, and high-tweet-attention groups. Higher attention levels are associated with progressively larger mean absolute next-day returns

Table 5 Tweet Attention and Magnitude of Next-Day Returns

Attention Level	N	Mean Tweets	Mean Absolute Next-Day Return (%)	Median Absolute Return (%)
Low	1,458	1.32	2.0543	1.4214
Medium	1,372	5.50	2.2141	1.5707
High	1,354	40.03	2.3813	1.6729

The results indicate a positive descriptive relationship between tweet attention and the magnitude of subsequent stock-price movements.

4.12 Tweet Volume and Absolute Next-Day Stock Returns

The relationship between tweet volume and the magnitude of subsequent stock-price movements was examined directly using the absolute value of next-day returns.

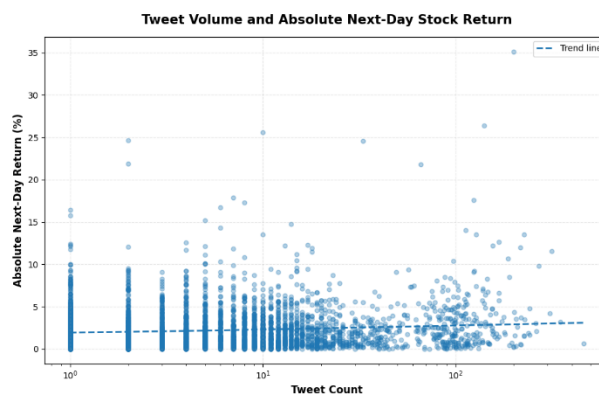


Figure 11. Tweet Volume and Absolute Next-Day Stock Return

The scatterplot presents tweet volume against the absolute value of the next-day stock return. The dashed line represents the fitted trend.

The Pearson correlation between tweet volume and absolute next-day return was 0.1406 ($p < 0.001$), while the Spearman correlation was 0.0636 ($p < 0.001$). Although the correlations are modest in magnitude, both indicate a statistically significant positive association.

The finding suggests that higher levels of social-media attention tend to coincide with greater subsequent stock-price movement, even though the direction of that movement cannot be inferred from attention alone.

4.13 Extreme Next-Day Price Movements by Tweet Attention Level

To further examine whether higher social-media attention is associated with unusually large stock-price movements, extreme next-day price movements were defined as observations for which the absolute next-day return was at least 5%. The proportion of extreme movements increased across the attention groups. Extreme movements occurred in 8.09% of observations in the low-attention group, compared with 9.99% in the medium-attention group and 11.37% in the high-attention group.

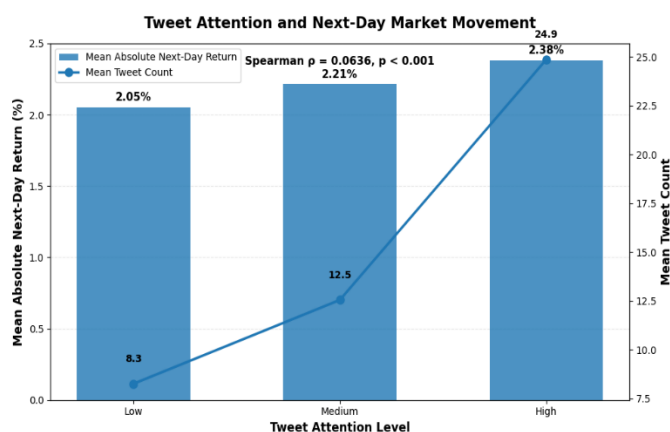


Figure 12. Extreme Next-Day Price Movements by Tweet Attention Level

The figure shows the percentage of observations with absolute next-day returns of at least 5% across low-, medium-, and high-tweet-attention levels.

The increasing proportion of extreme movements across attention levels provides additional descriptive evidence that higher social-media attention is associated with greater stock-price movement intensity. However, this comparison represents an association rather than a causal effect since heightened attention may also arise in response to underlying market or firm-specific events.

4.14 Stock-Wise Tweet Attention and Next-Day Market Movement

The relationship between tweet attention and absolute next-day returns was also examined separately for each stock using Spearman correlation.

The results showed considerable cross-sectional variation. The strongest positive association was observed for NOC ($\rho = 0.414$), followed by KO ($\rho = 0.257$) and NFLX ($\rho = 0.232$). Several stocks also showed statistically significant positive associations.

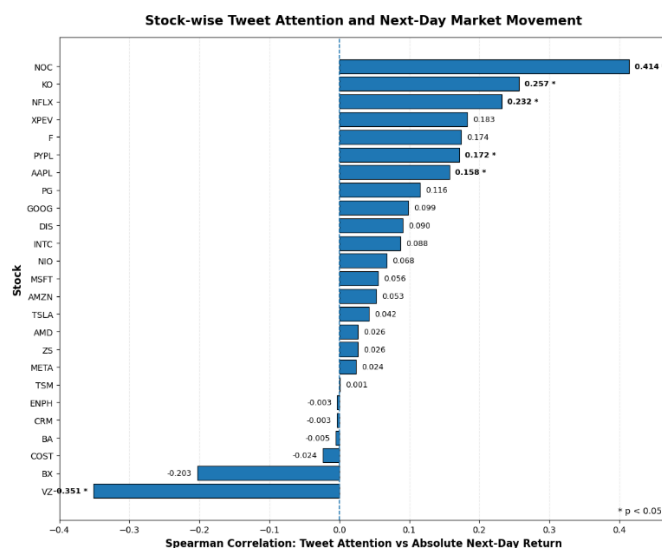


Figure 13. Stock-wise Tweet Attention and Next-Day Market Movement

The figure presents stock-wise Spearman correlations between tweet attention and absolute next-day stock returns. The asterisk indicates statistical significance at $p < 0.05$.

The stock-wise results indicate that the attention-movement relationship is not identical across all securities. Nevertheless, the presence of positive associations for several stocks supports the broader aggregate finding that greater social-media attention is associated with greater subsequent price-movement magnitude.

4.15 Controlled Regression of Next-Day Absolute Return

To determine whether the attention-movement relationship remained after accounting for market-related variables, a controlled regression model was estimated with absolute next-day return as the dependent variable.

The explanatory variables included log tweet count, lagged daily return, rolling volatility, log trading volume, and intraday range. Heteroskedasticity-consistent robust standard errors were used.

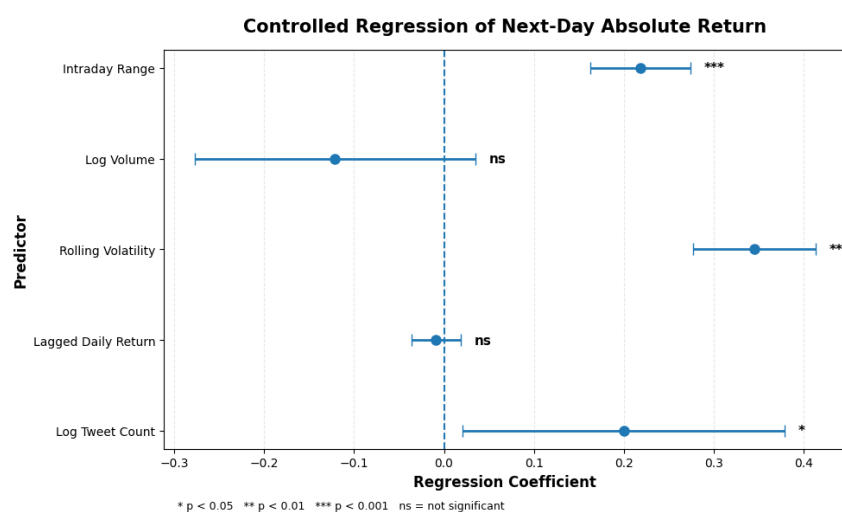


Figure 14. Controlled Regression of Next-Day Absolute Return

The figure presents regression coefficients and confidence intervals from the controlled regression model. Statistical significance is indicated as * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Table 6 Controlled Regression Results for Next-Day Absolute Return

Predictor	Coefficient	p-value	Significance
Log Tweet Count	0.1996	0.029	*
Lagged Daily Return	-0.0081	0.594	ns
Rolling Volatility	0.3446	<0.001	***
Log Volume	-0.1420	0.127	ns
Intraday Range	0.2248	<0.001	***
R ²	0.1484	—	—
Adjusted R ²	0.1473	—	—

The coefficient on log tweet count was positive and statistically significant ($\beta = 0.1996$, $p = 0.029$). Importantly, the association remained after controlling for recent return behaviour, volatility, trading volume, and intraday range.

Rolling volatility and intraday range were also strongly associated with subsequent absolute returns. In contrast, lagged return and log volume were not statistically significant in this specification.

The regression provides additional evidence that the positive association between social-media attention and the magnitude of subsequent stock-price movements is not explained solely by the bivariate relationship.

4.16 Classification Model Performance

Machine-learning classification models were employed to examine whether social-media sentiment and market variables could predict the direction of the next-day stock return.

A strict chronological train-test split was used. The training sample contained 5,025 observations, while the test sample contained 1,250 observations. The majority-class baseline accuracy on the test set was 55.12%.

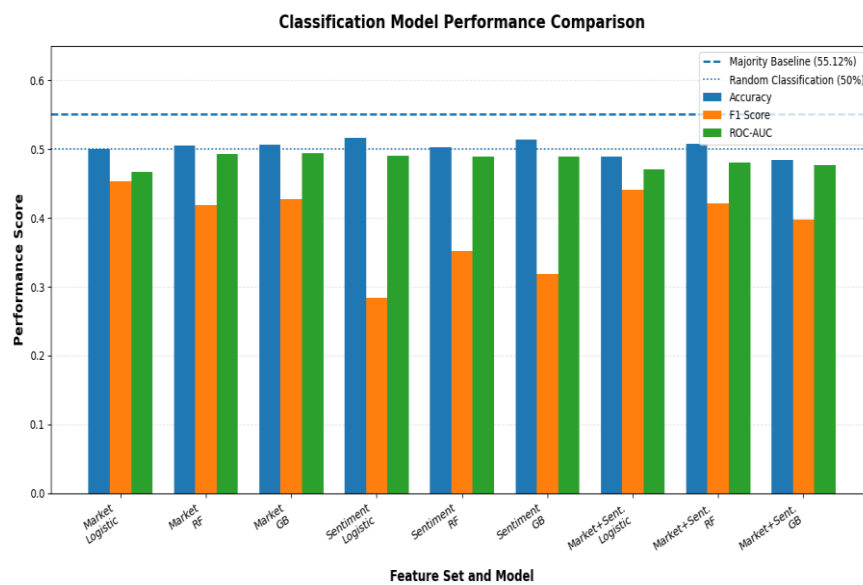


Figure 15. Classification Model Performance Comparison

The figure compares classification accuracy, F1-score, and ROC-AUC across market-only, sentiment-only, and market-plus-sentiment models. The horizontal reference lines indicate the majority-class and random-classification benchmarks.

Table 7 Classification Model Performance on the Chronological Test Set

Feature Set	Model	Accuracy	Balanced Accuracy	F1-Score	ROC-AUC
Market	Logistic Regression	0.5000	0.4964	0.4532	0.4672
	Random Forest	0.5056	0.4956	0.4192	0.4924
	Gradient Boosting	0.5064	0.4974	0.4271	0.4936
Sentiment	Logistic Regression	0.5160	0.4879	0.2840	0.4907

Feature Set	Model	Accuracy	Balanced Accuracy	F1-Score	ROC-AUC
	Random Forest	0.5024	0.4837	0.3521	0.4887
	Gradient Boosting	0.5144	0.4901	0.3187	0.4888
Market + Sentiment	Logistic Regression	0.4896	0.4857	0.4404	0.4704
	Random Forest	0.5072	0.4972	0.4211	0.4806
	Gradient Boosting	0.4848	0.4750	0.3981	0.4762

None of the classification models exceeded the 55.12% majority-class baseline. The highest accuracy among the reported models was 51.60%, obtained by the sentiment-only Logistic Regression model, which remained below the baseline. ROC-AUC values ranged approximately from 0.467 to 0.494, indicating performance close to or below random discrimination. The addition of sentiment variables therefore did not improve the ability of the models to predict next-day return direction.

4.17 Boost Next-Day Return Prediction Performance

XGBoost was additionally evaluated as a regression model for predicting the numerical value of the next-day return.

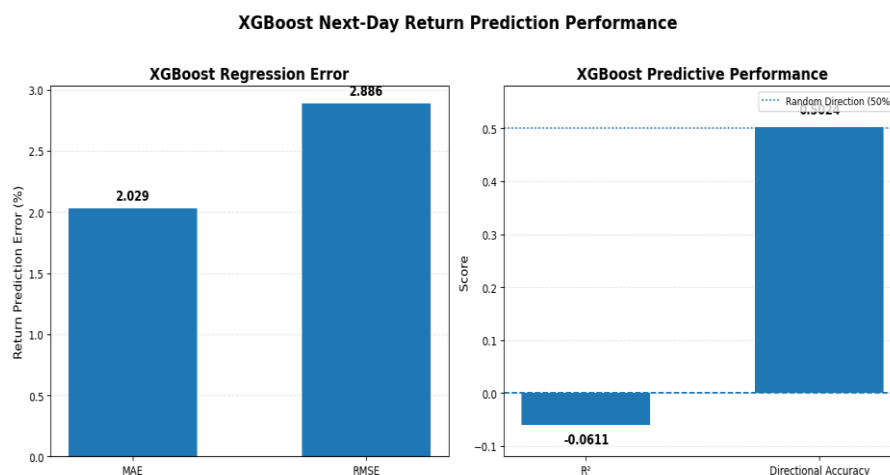


Figure 16. XGBoost Next-Day Return Prediction Performance

The figure presents the mean absolute error (MAE), root mean squared error (RMSE), R^2 , and directional accuracy obtained from the XGBoost next-day return prediction model.

Table 8 : XGBoost Next-Day Return Prediction Performance

Performance Measure	Result
MAE	2.029%
RMSE	2.886%
R ²	−0.0611
Directional Accuracy	50.24%

The XGBoost model produced an MAE of 2.029% and an RMSE of 2.886%. The negative R² of −0.0611 indicates that the model did not outperform a simple benchmark based on predicting the mean return. Directional accuracy was 50.24%, which is close to random classification.

Therefore, the machine-learning regression results do not indicate meaningful predictive performance for next-day returns.

4.18 Overall Findings

The empirical results reveal a clear distinction between social-media sentiment and social-media attention.

First, sentiment showed weak and inconsistent relationships with next-day stock returns. The overall correlation was close to zero, stock-wise correlations were small, lagged relationships remained weak, and classification models failed to outperform the majority-class benchmark.

Second, tweet volume showed a positive relationship with the magnitude of subsequent stock-price movements. Mean absolute returns increased from 2.05% in the low-attention group to 2.38% in the high-attention group. The Pearson correlation between tweet volume and absolute next-day return was 0.1406, while the controlled regression produced a positive and statistically significant coefficient for log tweet count.

Thus, the central empirical pattern of the study is that social-media attention appears more informative about the intensity of subsequent stock-price movements than about their direction.

4.19 Hypothesis-Wise Findings

H1: Social-media sentiment is significantly associated with subsequent stock-market returns and the direction of next-day stock-price movements.

The results do not provide sufficient evidence to support H1. The overall sentiment-return relationship was weak, the lagged correlations were close to zero, and the machine-learning classification models failed to outperform the majority-class benchmark.

Decision: H1 is not supported.

H2: Social-media attention, measured by tweet volume, is significantly positively associated with the magnitude of subsequent stock-price movements.

The results provide evidence supporting H2. Tweet volume was positively associated with absolute next-day returns, the high-attention group experienced greater average absolute movements than the low-attention group, and the controlled regression showed a statistically significant positive coefficient for log tweet count.

Decision: H2 is supported.

Table 9 Summary of Hypothesis Testing

Hypothesis	Main Evidence	Decision
H1: Sentiment is significantly associated with subsequent return direction	Weak correlations; classification accuracy below 55.12% baseline; ROC-AUC near 0.50	Not Supported
H2: Tweet attention is positively associated with return magnitude	Pearson $r = 0.1406$; high attention means absolute return = 2.38%; regression $\beta = 0.1996$, $p = 0.029$	Supported

4.20 Summary of Results

The results can be summarized in four major findings.

1. Social-media sentiment did not demonstrate a strong or consistent relationship with next-day stock returns. The overall correlation was exceedingly small, and stock-wise relationships were weak.
2. Sentiment did not provide reliable directional prediction. Classification models incorporating sentiment variables failed to exceed the majority-class baseline, while XGBoost also showed weak predictive performance.
3. Social-media attention was positively associated with the magnitude of next-day stock-price movements. Higher tweet activity corresponded to higher mean absolute returns and a greater frequency of extreme price movements.
4. The attention-movement relationship remained statistically significant after controlling for market variables. The controlled regression showed a positive coefficient for log tweet count, indicating that the relationship was not eliminated by accounting for recent returns, volatility, trading volume, and intraday price range.

Overall, the findings suggest that social-media activity may be more useful as an indicator of potential market movement intensity than as a reliable predictor of market direction. This distinction forms the central empirical contribution of the study.

Chapter 5: Discussion

5.1 Introduction

The present study investigated the association between social-media sentiment, social-media attention, and stock-market behaviour across 25 stocks. The analysis considered both the direction of next-day stock-price movements and the magnitude of these movements. The findings indicate that social-media sentiment did not demonstrate a strong or consistent relationship with subsequent stock returns or next-day price direction. In contrast, social-media attention, measured through tweet volume, showed a positive and statistically significant association with the magnitude of subsequent stock-price movements.

The findings therefore highlight an important distinction between sentiment and attention. Sentiment provided little evidence of a reliable directional relationship with subsequent stock returns, whereas higher levels of social-media attention were associated with larger absolute next-day stock-price movements. This distinction is central to interpreting the empirical results of the study.

5.2 Social-Media Sentiment and Stock Returns

The empirical results, however, do not suggest a strong and consistent relationship between social-media sentiment and the subsequent stock returns. Both the Pearson and Spearman correlations between sentiment and next day return were small and statistically insignificant. The results of the stock-wise analysis indicated that one stock (NIO) had a statistically significant Pearson relationship, and the relationship was negative.

This is corroborated by the lagged analysis. There was only a very weak relationship between sentiment measured one day in advance and subsequent returns, while the relationships across longer horizons were unimportant. The Spearman correlation was significant for the 2-day period but was exceptionally low and of little practical value as evidence for predictive value.

These findings are confirmed by the machine-learning classification results. No one of the classification models consistently did better than the baseline majority class on the chronological test set. This indicates that the sentiment and market characteristics employed in this study did not have enough information to make a reliable prediction of the price direction for the next day.

This evidence is aligned with the notion that the sentiment of investors might have an effect on the market under specific conditions, but it does not necessarily constitute a reliable direction signal for stocks and time periods (Baker & Wurgler, 2007; Sul et al., 2017; Behrendt & Schmidt, 2018).

5.3 Social-Media Attention and Stock-Market Movements

This was no longer the case when social-media attention was analysed. The absolute value of next-day returns was positively and statistically significantly correlated with Tweet volume. The Pearson correlation was ~ 0.141 , and the Spearman correlation was ~ 0.064 , both of which were statistically significant.

The attention-group analysis further revealed that stocks with more tweet activity had larger average absolute returns on the day following a tweet. In the low-attention group, there was a mean absolute return of about 2.05%, whereas the mean absolute return in the high-attention group was 2.38%. The difference of ~ 0.33 percentage points was statistically significant using the Welch test.

Important, this is a relationship about the magnitude of price moves not the direction of the price moves. There was no indication of price direction as attention was higher. Rather, subsequent price moves were greater during periods of increased attention.

Further evidence is given by the controlled regression analysis. Log tweet volume was positively and significantly related to the absolute next-day return after other factors, such as previous returns, historical volatility, trading volume and intraday price range, were controlled for. These results were also robust to the inclusion of stock fixed effects and stock-clustered inference. This finding is in line with studies that highlight the importance of the investor attention as a relevant market mechanism. There are several studies that have demonstrated a relationship between attention measures and trading activity, information diffusion, and stock-market volatility (Da et al., 2011; Andrei & Hasler, 2015; Audrino et al., 2020; Rakowski et al., 2021).

5.4 Comparison with Previous Studies

The results do not always corroborate the previous studies on social media and financial markets. Previous research has found that in specific circumstances, Twitter sentiment can include data on the return or the movement of a stock (Bollen et al., 2011; Sprenger et al., 2014; Sul et al., 2017). Other studies, however, have emphasized

the significance of Twitter activity and investor attention besides the sentiment alone (Behrendt & Schmidt, 2018; Rakowski et al., 2021).

This study contributes to this literature by disaggregating the sentiment and disaggregating the direction of return from the magnitude of return. This separation can be used to explain why sentiment-based prediction can sometimes result in sub-optimal predictions, even though social-media activity is sometimes correlated with future market turbulence.

The results are also in line with the existing literature on investor attention, which indicates that heightened attention may affect the processing of investor information and the way in which this information is reflected in asset prices (Barber & Odean, 2008; Hirshleifer & Teoh, 2003).

5.5 Theoretical and Practical Implications

The results, in theory, can be used to better segment social-media information in financial markets. Social media should not necessarily be viewed only as a source of positive or negative sentiment. The level of "noise" around a stock can be an indicator of how much market activity is occurring around that stock.

In practice, investors and market analysts might thus discover that the volume of tweets is a valuable measure of attention or risk monitoring, but sentiment alone has proven to be a more complex buy-or-sell signal. A rise in the amount of social media activity could mean that there may be more significant price movements in the future, but it does not predict the direction in which they will move.

The observed relationship is important to note as it is not necessarily causal. While there might be greater activity in the market due to more attention in the social media, there could also be more attention in the social media because there is more activity in the market. The observational/lagged design helps to correct some timings issues but is not causal.

5.6 Limitations of the Study

There are several limitations on the study. Firstly, the analysis is over a brief time horizon using only 25 stocks, limiting the scope of the results. Second, the analysis relies on Twitter activity, meaning it does not account for attention garnered outside of Twitter, on other social-media platforms or financial information outlets.

Third, the sentiment measures derived from the information in the text may not be a perfect reflection of sarcasm, context, or more complicated investor opinions. Fourth, although lagged variables and control variables were used, the observational design cannot eliminate endogeneity or reverse causality. Lastly, stock returns are not normally distributed and there are extreme observations which might impact some statistical relationships even though using robust inference.

5.7 Future Research

Future studies might be conducted on longer time horizons, more stocks, and various social-media platforms. Future research might also be done for intraday relationship to check if social-media attention is responding or preceding short-term price changes.

More sophisticated causal designs may be able to tell if more attention really does lead to greater subsequent volatility. Future studies may also involve a comparison of different measures of attention, sentiment intensity, investor disagreement, and spread of information.

Chapter 6: Conclusion And Recommendations

6.1 Introduction

This chapter summarizes the overall findings of the study and draws conclusions from the empirical findings discussed in the previous chapters. It provides an overview of the research findings regarding the research

hypotheses and emphasizes key contribution of the study. It gives suggestions and directions for further research as well.

The study explores the relationship between social-media sentiment and social-media attention data with the stock-market behaviour of the 25 stocks for the duration of the study. Descriptive statistics, correlation analysis, lagged relationships, attention-group comparisons, regression models, fixed-effects analysis, and machine-learning models were all considered.

The study considered two facets of stock-market behaviour: the direction of the next day return and the size of the next day price move. The distinction is significant because while social-media activity can have no impact on whether a particular stock will rise or fall, it could be correlated with the size of the subsequent price swing.

6.2 Summary of the Study

For 25 selected stocks, the study examined the correlation between sentiment in the social media, social media attention, and stock market behaviour. Sentiment on social-media was analyzed with respect to returns on the stock and direction of next-day price movements, while social-media attention was measured by the number of tweets and analyzed with respect to the size of subsequent stock-price movements.

Multiple statistical and machine learning methods were used in the study. These encompassed correlation analysis, lagged analysis, sentiment-intensity analysis, attention-group comparisons, controlled regression, stock fixed-effects analysis, and machine-learning models for next-day return-direction analysis and return prediction.

The analysis was developed to differentiate between two distinct functions of social-media information. The first is whether the attitude in social media activity will offer any clues as to the path of future stock price changes. The second is if there is any correlation between the amount of attention that a share receives and the size of the price action.

6.3 Major Findings

The main results of the study are presented below:

1. There was weak and inconsistent association between sentiment in social media and next day stock returns. Overall sentiment–return relationship was low and stock-specific correlations were weak. One stock had a Pearson correlation coefficient that was statistically significant at 5%.
2. There was no consistent relationship between lagged social-media sentiment and future stock returns. The relationships were weak and did not show a definite trend of predictive value throughout the various future-return horizons studied.
3. Sentiment on social media was not a reliable indicator of directionality. The machine-learning classification models, such as Random Forest, Logistic Regression, Decision Tree, and Gradient Boosting did not perform better than the majority class baseline of 55.12% on the chronological test set. This means that it was not possible to reliably predict stock-price direction using the variables employed in the classification models in the current data set.
4. The size of the stock-price moves was positively related to social-media attention. Absolute next day returns were positively related to tweet volume. The Pearson correlation coefficient for tweet volume and the absolute next day return was 0.1406, and the Spearman correlation coefficient between tweet volume and absolute next day return was 0.0636.
5. There is a positive correlation between higher-attention observations and the absolute next-day returns. The average absolute next-day returns went up from about 2.05% in the low-attention group to 2.21% in the medium-attention group to 2.38% in the high-attention group. The Welch test showed the difference between the low- and high-attention group was statistically significant.
6. After adjustment for pertinent market parameters, the attention–movement relationship was statistically significant. Although the controlled regression analysis revealed a positive coefficient for log tweet

count, it did so while controlling for prior returns, historical volatility, trading volume, and intraday price range, and the results were statistically significant.

The positive attention–movement association was robust to stock fixed effects. This means that the relationship observed remained when the persistent differences between the 25 stocks were removed. The finding should be understood as an association, however, and not as causal.

Overall, the results suggest there is a difference between social-media sentiment and social-media attention. Higher levels of sentiment had no clear association with the magnitude of subsequent stock-price movements, while more social-media attention was associated with greater subsequent stock-price movements.

6.4 Conclusions based on the hypotheses

Hypothesis 1

H2: The sentiment of social media is strongly linked to the direction of next-day stock-price changes and subsequently, in-sample stock-market returns.

The result of the test fails to support H1. The overall sentiment–return relationship was not strong, the stock-wise correlations were typically small, and sentiment in each time was not consistently linked to return in the following period. Furthermore, the machine learning classifies models could not beat the majority class next-day return direction benchmark.

Hence H1 is not accepted.

Hypothesis 2

H2: Subsequent stock-price movements (measured by absolute next-day stock returns) are significantly positively related to social-media attention (measured by tweet volume).

The results present evidence in support of H2. Tweet volume was positively associated to the absolute next day returns and observations having higher tweet attention had larger mean absolute next day returns. The same positive coefficient was found for log tweet count in the controlled regression analysis and was statistically significant. The attention–movement relationship was also statistically significant for the stock fixed-effects specification.

Hence, H2 is correct.

6.5 Key Contribution of the Study

One of the study's contributions is the separation of sentiment and attention, as it relates to stock market behaviour, in social media.

The results show that sentiment was not a good predictor of the direction of next-day stock-price moves. Social-media attention, however, was statistically significantly positively correlated with the size of future stock-price changes.

The results indicated that the level of social-media attention might be more useful to infer the magnitude of the subsequent stock price moves than the direction. This offers a more nuanced view of the impact of social-media activity in financial markets.

However, the results must be viewed as evidence of association and not causation. Lagged variables, market controls, and stock fixed effects were also added to the analysis, but the observational design does not imply that increased tweet activity leads to increased subsequent price movements.

6.6 Recommendations

Based on the results of the study, the following recommendations are made:

1. For investors: social-media sentiment should not only be used as an indicator of next-day stock-price direction because of the weak relationships and classification models that do not outperform a majority-class benchmark.
2. For market analysts: The volume of tweets could serve as an attention or risk-monitoring factor when predicting the probability of the next big stock-price move. It should be used in conjunction with other market information and not used solely as an indication for trading.
3. For researchers: further research is needed to assess if the attention-movement relationship persists in the future under longer time horizons, larger sample of stocks and social-media data from other platforms.
4. Methodological research: Future studies can use intraday data, other proxies of investor attention and more rigorous causal-identification techniques to gain a more nuanced understanding of the time lag between social-media activity and stock-market changes.

6.7 Future Research Directions

This study suggests a number of directions for further research:

First, future studies could be extended in the number of stocks studied as well as in the length of the observation period. A longer time frame would enable the researchers to determine if the social-media attention-stock-price relationship varies under different market conditions.

Second, it is possible to study the variation in industries, market capitalisation groups and in high market stress and low market stress periods in future. Such analysis can be used to determine if the attention–movement relationship is stronger for specific kinds of stocks or in specific market conditions.

Third, future studies might add other measures of social-media attention, like engagement, likes, replies, reposts, and user level activity, as well as tweet volume.

Fourth, intraday data may elicit a better understanding of the timing of the relationship between social-media activity and the market movements. This could be useful for identifying if shifts in social-media focus drive market moves or if they are simply a response to the information becoming known to investors first.

Lastly, future research might use more powerful causal-identification techniques to determine if social media attention per se is linked to subsequent stock price changes or if the link is due to a common omitted information or market event.

6.8 Final Conclusion

The research results indicate that social-media sentiment in the data set examined in this study is not a good predictor of the direction of stock-price changes the following day. Empirical results indicate that sentiment is only weakly and inconsistently related to future stock returns, and the machine learning classification models did not perform better than the majority class benchmark.

By contrast, the number of tweets was positively and significantly associated with the size of the subsequent stock-price changes. This effect was statistically significant and, importantly, larger absolute next-day returns were associated with higher levels of tweet activity.

The crux is consequently not that social media offers an effective way to predict the direction of a stock price. Instead, the results indicate that the amount of social-media coverage can have information on the magnitude of future stock-price changes. This sentiment-attention dichotomy offers a more nuanced view of the social media in financial markets.

But these findings are associations and not causes. As the study is observational, it cannot say whether the social-media attention leads to bigger stock-price moves or whether the two are driven by other information or market factors.

Overall, the study adds to the literature by showing that the informational content of social media could be different depending upon which dimension is analysed: sentiment is more like pointing a direction and attention is more like predicting the magnitude of future market moves.

Final Reference List — APA 7th Edition

- [1] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1), 43–76. [https://doi.org/10.1016/S0304-405X\(01\)00055-1](https://doi.org/10.1016/S0304-405X(01)00055-1)
- [2] Andrei, D., & Hasler, M. (2015). Investor attention and stock market volatility. *The Review of Financial Studies*, 28(1), 33–72. <https://doi.org/10.1093/rfs/hhu059>
- [3] Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance*, 59(3), 1259–1294. <https://doi.org/10.1111/j.1540-6261.2004.00662.x>
- [4] Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, 36(2), 334–357. <https://doi.org/10.1016/j.ijforecast.2019.05.010>
- [5] Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4), 1645–1680. <https://doi.org/10.1111/j.1540-6261.2006.00885.x>
- [6] Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2), 129–151. <https://doi.org/10.1257/jep.21.2.129>
- [7] Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behaviour of individual and institutional investors. *The Review of Financial Studies*, 21(2), 785–818. <https://doi.org/10.1093/rfs/hhm079>
- [8] Behrendt, S., & Schmidt, A. (2018). The Twitter myth revisited: Intraday investor sentiment, Twitter activity, and individual-level stock return volatility. *Journal of Banking & Finance*, 96, 355–367. <https://doi.org/10.1016/j.jbankfin.2018.09.016>
- [9] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- [10] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- [11] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- [12] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [13] Bukovina, J. (2016). Social media big data and capital markets—An overview. *Journal of Behavioural and Experimental Finance*, 11, 18–26. <https://doi.org/10.1016/j.jbef.2016.06.002>
- [14] Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372. <https://doi.org/10.3368/jhr.50.2.317>

- [15] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- [16] Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2024). The social signal. *Journal of Financial Economics*, 158, 103870. <https://doi.org/10.1016/j.jfineco.2024.103870>
- [17] Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5), 1461–1499. <https://doi.org/10.1111/j.1540-6261.2011.01679.x>
- [18] Edmans, A., García, D., & Norli, Ø. (2007). Sports sentiment and stock returns. *The Journal of Finance*, 62(4), 1967–1998. <https://doi.org/10.1111/j.1540-6261.2007.01262.x>
- [19] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. <https://doi.org/10.2307/1912773>
- [20] Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance*, 64(5), 2023–2052. <https://doi.org/10.1111/j.1540-6261.2009.01493.x>
- [21] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. <https://doi.org/10.2307/2325486>
- [22] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
- [23] Harvey, C. R., & Liu, Y. (2016). Lucky factors. *Journal of Financial Economics*, 121(2), 306–327. <https://doi.org/10.1016/j.jfineco.2015.09.005>
- [24] Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1–3), 337–386. <https://doi.org/10.1016/j.jacceco.2003.10.002>
- [25] Rakowski, D., Shirley, S. E., & Stark, J. R. (2021). Twitter activity, investor attention, and the diffusion of information. *Financial Management*, 50(1), 3–46. <https://doi.org/10.1111/fima.12307>
- [26] Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5), 926–957. <https://doi.org/10.1111/j.1468-036X.2013.12007.x>
- [27] Sul, H. K., Dennis, A. R., & Yuan, L. (2017). Trading on Twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 48(3), 454–488. <https://doi.org/10.1111/deci.12229>
- [28] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>
- [29] Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437–1467. <https://doi.org/10.1111/j.1540-6261.2008.01362.x>
- [30] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838. <https://doi.org/10.2307/1912934>
- [31] *The causal relationship between social media sentiment and stock return: Experimental evidence from an online message forum.* (2022). *Economics Letters*, 216, 110598. <https://doi.org/10.1016/j.econlet.2022.110598>
- [32] Wu, H. (2019). Social media attention and stock returns. *International Review of Financial Analysis*, 65, 101364. <https://doi.org/10.1016/j.irfa.2019.101364>