ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

International bachelor's in Economics & Business Economics

Major in Financial Economics

Can Twitter data regarding popular technology stocks improve the predictive power of Fama & French asset pricing models?

Author: Vid Tominec

Student number: 560361

Email: 560361vt@eur.nl

Thesis supervisor: Dr. Ruben de Bliek

Second assessor: Dr. Laurens Swinkels

Finish date: 20/09/2023

Abstract

This study aims to explore whether Twitter sentiment can be used to improve the predictive efficiency of the Fama & French 3 and 5-factor models for a portfolio comprised of Tesla, Apple, Amazon, Google and Microsoft, for a period between 2015 and 2020. More than 4 million tweets regarding the companies are analyzed using VADER, a social media specific natural language processing tool, to expand both ordinary least squares models with a sentiment component. The sentiment analysis was found to be insignificant and ineffective at increasing the explanatory power for either of the models. In addition, the 5-factor model was shown to be a better fit for the sample in explaining excess returns. The study questions the validity of using social media as a reliable source of information regarding investor sentiment and its applicability in asset pricing models.

Table Of Contents

1.Introduction

Measuring portfolio excess returns and trying to understand which factors affect them has been a common research topic in finance for a long time. One of the earliest models for modelling excess returns is known as the "Capital Asset Pricing Model" (CAPM), developed by William F. Sharpe (1964), which models a linear relationship between excess market returns and risk. Arguably the most influential asset pricing models have been developed by Fama & French in the last 30 years, with their 3 and 5-factor models. Both models attempt to explain portfolio performance using firm and portfolio-specific variables that attempt to establish a statistically significant relationship with excess returns. Here, the 5-factor model will be expanded upon with an additional variable which serves as a proxy for investor sentiment regarding a stock portfolio of stocks with a heavy media presence. This variable seeks to improve the model's explanatory power of excess returns by trying to capture the effect of investor behavior regarding stock performance between 2015 and 2020.

Investors' perceptions and public opinion regarding both recent financial news and information have long been important aspects of stock investing and portfolio formation. Ever since the introduction of the internet, both news and social media have played a significant role as a facilitator for investors, both retail and institutional, to exchange information and opinions. With quicker reactions and more weight put on public sentiment regarding the market, investors are increasingly more reliant on information gathered from the internet. Social media popularity of brands has also been shown to be statistically correlated to the underlying stock price, suggesting the presence of information bias and herd behavior (O'Connor, 2013). Previous literature has tried incorporating sentiment scores into different asset pricing methods to see their explanatory power. Sentiment with regards to tradable assets is analyzed based on user-generated data on social media platforms regarding either a technical or fundamental aspect of a security to establish a score capturing recent trends in herd opinion (Wankhade, Rao & Kulkarni, 2022). Incorporating sentiment scores by analyzing social media opinions and predictions has been shown to increase the predictive power of Fama, French and Carhart models (Houlihan & Creamer, 2017). While Houlihan & Creamer (2017) analyzed over 5000 companies using roughly 4 million tweets, the focus here will be using a similar sized tweet database regarding only 5 technology companies, over a longer time frame.

The focus of this paper is to explore the effects of investor sentiment from "Twitter", a short-form broadcasting social media platform, on the predictive performance of the Fama & French 3 and 5-factor models on a stock portfolio consisting of the following five companies: *Google, Apple, Amazon, Tesla* and *Microsoft*. When it comes to media presence, the technology sector is typically the one that receives the most media attention and follow-up social media discussion (Gandhi, et al., 2020). Twitter, due to its "micro-blogging" nature of short, highly time relevant pieces of shared information make for a suitable platform for financial discussion to take place. Sentiment scores derived from Twitter data on a daily time frame have been attributed to an 86.7% accuracy in predicting closing values of the "Dow Jones Industrial Average" (Bollen, Mao & Zheng, 2011). News sentiment scores have likewise been shown to have a statistically significant and positive relationship with S&P500 price movements (Costola et al., 2023). Most literature insofar has been geared towards the market-wide effects of news and social media sentiment. Here, a closer look at returns specific to the five companies will be examined with regards to sentiment analysis, to examine how relevant the role of investor sentiment is in portfolios consisting of stocks with some of the highest media coverage.

The sample will consist of five publicly traded US companies in the technology sector, namely *Tesla, Apple, Google, Facebook* and *Amazon*, for a time between 2015 and January 2020. Using this approach, stocks with a large media presence are selected, providing a consistent and large amount of text for the sentiment score analysis. Data regarding the stocks and the market will be collected through *Yahoo Finance*. To construct a sentiment variable that will be added to the asset pricing models, Natural Language Processing (NLP) will be used in combination with Twitter data regarding the relevant stocks in the portfolio. The main approach implemented to construct the sentiment variable is known as *VADER,* short for *Valence Aware Dictionary and sEntiment Reasoner,* a sentiment analysis method specifically designed to analyze user generated content within the context of social media (Hutto & Gilbert, 2014). Using this, a normalized sentiment variable between 0 and 1 can be constructed for each security analyzed, for any given month, providing a proxy for sentiment during that time frame. Hence, we arrive at the following research question:

What is the impact of investor sentiment derived from Twitter data on the predictive performance of the Fama & French 3 and 5-factor models for a portfolio consisting of Tesla, Apple, Google, Facebook, and Amazon between 2015 and January 2020?

The expectation for the study is hard to predict due to high levels of noise in Twitter data as well as classifier accuracy likely not being perfect. Given that the Fama and French models can capture a relatively high amount of variation in portfolio excess returns, the improvement of these models with the addition of a sentiment variable is likely to be small. While previous research has successfully implemented such additions to asset pricing models, the main issue with a sentiment score is likely to be the quality of the data and ensuring that data is only relevant to the respective portfolios and securities. In addition, the idea of exploring the role of investor sentiment in portfolio performance is generally speaking a difficult task due to the very high levels of noise and random variance that come from the use of social media. For instance, the Natural Language Processing model presented above does not account for sarcasm, which is likely to affect their predictive power. A causal relationship is also hard to establish in such cases since factors such as seasonality, institutional behavior and index rebalancing can lead to potential cases of spurious correlation between sentiment and excess returns.

The rest of the paper is structured as follows. The theoretical framework discusses the theory behind the models used to explain excess returns, as well as the academic link between investor sentiment and asset pricing. Following, the methodology section, explains the statical method behind the Fama & French models, as well as the sentiment analysis methodology, in addition to the construction of the models. In the result section, tables with the relevant model results are discussed, as well as the interpretations of the models. The last two sections are a discussion, where the economic significance of the results in congruence with the literature presented are discussed, as well as a conclusion summarizing the research.

2. Theoretical Framework

2.1 Excess Return Models

Excess returns in the stock market are what investors seek when it comes to picking and choosing the stocks which comprise their individual portfolio, in hopes of outperforming the stock market. In general, a portfolio exhibits excess returns over a period when the appreciation of its assets is higher than that of the general stock market, or a different measurement proxy. The typical understanding of excess returns in investing is that an investor is likely to be rewarded with excess returns when one takes on more risk (Glosten et al., 1993). Understanding the relationship between risk and return has been at the center of the discussion on which assets are likely to experience excess returns, given a specified timeframe. Some of the earliest academic research has proposed a linear relationship between risk and excess returns, capturing the relationship using a model known as Capital Asset Pricing Model, also known as CAPM, coined by Sharpe (1964). The CAPM model suggests that if a specific security is more volatile than the general market, the investors should be compensated for taking on more systematic risk with higher returns. This general premise of the model, however, has been highly debated in recent academic history, with numerous studies concluding that systematic risk alone cannot explain excess returns fully (Galagedera, 2007). For instance, the average returns of companies with a smaller valuation have been found to be higher than the predicted returns of the CAPM model, as well as companies with a higher book-to-market ratio (Banz, 1981, and Statman, 1980, as cited in Elbannan, 2014). In addition, research on risk factors and their evolutions suggests that due to their ever-evolving nature, it might be unreasonable to expect that static, factor specific models will be able to accurately explain the risk-return relationship, in the long run (Maiti, 2020). Despite this, for practical reasons, different factor models that can explain a good portion of excess returns will most likely still be used, either as a benchmark or as a tool.

Later models have tried to incorporate additional explanatory variables to explain excess returns of securities. The three-factor model expands the CAPM by adding size and value as explanatory variables which help to explain excess returns, in addition to systematic risk. The reasoning behind the two additional variables in the model, is due to the empirical findings that firms with a smaller

market capitalization and a higher book-to-market ratio tend to perform better (Fama and French, 1992). The three-factor model is seen as the superior model with a higher explanatory power of excess returns in most cases, when compared to the CAPM model (Blanco, 2012). While the model was later expanded to the five-factor model in 2015, accounting for investments and profitability characteristics, the statistical improvement over the three-factor model has not been consistent throughout literature. Showcased by Jiao & Lilti (2017), the five-factor model has performed better in the US stock market, as compared to the Chinese stock market. Other research has found similar findings when applied to other markets, suggesting that the explanatory power of both the 3 and 5 factor models tend to vary across different financial markets (Griffin, 2002).

In addition, practical application is easier with the three-factor model due to data availability constraint when constructing the additional variables in the five-factor model (Fama & French, 2015). Here, for the purpose of the study, both models will be used and analyzed, to see the difference between their respective performances. Both factors expand on the CAPM model with additional risk-factors, with the 3-factor model adding *size* and *value*, and the 5-factor model adding *investment* and *profitability* in addition. The *size* factor captures risk associated with the notion that stocks with smaller market capitalizations tend to outperform stocks with larger market capitalizations, while the *value* factor captures risk based on the argument that stock with low book-to-market ratios (referred to as *value* stocks) tend to outperform stocks with high book-to-market ratios (referred to as *growth* stocks) (Fama & French, 1992). The additional factor of *profitability* in the 5-factor model captures risk with the notion that firms with higher levels of profitability tend to outperform those with lower levels of profitability, with the *investment* factor arguing that firms which invest more aggressively tend to outperform firms that invest more conservatively (Fama & French, 2015). The specifics of the constructions of all these parameters are discussed in the following section of the paper. In addition, the focus on technology stocks here is due to the social media component, since previous research has shown that Twitter data can be especially useful in analyzing technology stocks, hence a similar approach is used here to see if the models can be expanded upon (Vu et al. 2012). Given that the 5-factor model is simply an extension of the 3-factor model and should capture more of the variance in excess returns, we explore the following hypothesis pertaining to given set of companies:

H1: The 5-factor Fama & French model will explain a higher share of the variance in excess returns, on a portfolio level, for selected tech stocks

2.2 Investor Sentiment and social media

A long-standing theory developed in part by Fama (1970), called the efficient market hypothesis (EMT) postulates the notion that prices of tradable securities on the market reflect certain information about them. The weakest form of EMT suggests that current prices reflect all past prices, whereas the strongest form suggests that all known information, both public and private, is reflected in the current price of a security. The implication of this theory suggests that earning excess returns without taking on more risk by taking advantage of mispricing should not be possible in most cases. Literature such as the three-factor model (Fama and French, 1992), as well as the momentum effect (Jagadeesh and Titman, 1993) have documented exceptions where pricing can either be explained through size and valuation factors, or through momentum, suggesting that investors can exploit such anomalies to outperform the market. In addition to market anomalies, EMT has been a debatable topic in financial research due to it disregarding the presence of numerous biases that are present in retail investors, due to the underlying assumption of rationality (Lo, 2007).

While the idea and the presence of "investor sentiment" has numerous interpretations, a common framework of thinking about sentiment is the *market participants' beliefs regarding future cash flows of an underlying asset* (Zhang, 2008). A more practical meaning implies that investor sentiment is essentially one's belief about what a company's intrinsic value should be, given the information available and its interpretation. Research in behavior finance has tried to investigate the underlying human behavior traits that lead to changes in investor sentiment and beliefs about market movements and company valuations. Early research has shown that investors tend to conform to the "representative bias", which is the tendency to judge the probabilities of certain outcomes based on the outcomes of similar past events (Kahneman & Tversky, 1972). Combined with "availability bias", which is the notion of relying on easily available or dramatic information

(Tversky & Kahneman, 1973), it becomes clearer why mispricing in the market is present. This can manifest itself in ways such as investors overestimating the impact of recent news on the stock market, as well as the practice of comparing events of economic downturn to each other and drawing conclusions. For instance, investors are more likely to invest in industries in which they work, as well as companies which are more likely to match their preferred characteristic (Pompian, 2006, as cited in Leković, 2020). In addition, behavioral finance has argued against the efficient market hypothesis using the concept of herd behavior, which is the act of ignoring private information in favor of public opinion (Banerjee, 1992). This phenomenon can lead to a disparity between the actual valuation of an asset or a market and its fundamental value, leading to so called "bubbles" in the market, a situation where asset classes are significantly overvalued.

Prior to the onset and the widespread use of social media, researchers in behavior finance have observed biases and psychological phenomenon that have led to irrational decision making in the market. With the recent rise of social media, studies have shown that it can exacerbate the presence of herd behavior in the stock market (Li et al., 2023). Li et al. also found that herd behavior is more likely to be present in individual investors, when compared to informed, institutional investors (2023). This suggests that investors who look for investing advice on social media are more likely to be influenced by the sentiment of similar opinions, thus giving social media sentiment predictive power in explaining stock market returns. With numerous social media platforms serving as places to form communities where investing related topics are discussed, Twitter, a short-form social media platform, has prevailed as the most popular one for such purposes (Shiva & Singh, 2020). For instance, investigations between the market sentiment extracted from Twitter and short-term movements of the NASDAQ-100 have found the predictive power of sentiment scores to be upwards of 88% (Rao & Srivastava, 2012).

2.3 Sentiment Score

To apply the concept of investor sentiment to the quantitative method of asset pricing, namely the Fama & French 5 factor model, a so called "sentiment score" is required, to bridge social media content and asset pricing components. The need for a sentiment score stems from the fact that

despite social media influencing investors' beliefs about the market, to see its effects one must quantify sentiment prior to its use in a statistical model. A sentiment score, depending on the machine learning technique used, is typically a score ranging between –1 and 1, with sentiment going from extremely negative (-1) to extremely positive (+1), with 0 being neutral. To construct a sentiment score, a social media platform needs to be used to gather a sufficient dataset of user generated content regarding the securities being analyzed, to capture the general sentiment of the security being analyzed. Prior research has shown that Twitter remains one of the most popular social media platforms for investors choosing where to source information regarding stock when forming portfolios (Shiva & Singh, 2020). In addition, the nature of Twitter's *micro-blogging* format, short form, highly relevant information has been shown to be preferred over longer form financial news that take longer to read in recent years, making its content a good proxy for investor sentiment (Corea, 2016).

When it comes to the specific application of Twitter data, the matter becomes slightly more intricate. While certain studies have shown that news sentiment is correlated to the movement of the stock market, both in returns and volume, conflicting research has shown that other non-financial stock market related data, such as CEO Twitter publications have very little to no statical effect on price performance (Smith, 2022). The reasonable explanation of what is going on here is that most likely, due to a high level of noise and limitations of using sentiment analysis tools for analyzing market or companywide specific beliefs, it is difficult to establish a correlation between market movements and social media content. Here, the exploration is slightly different, with the goal of the research seeking an improvement to asset pricing models which have been shown to capture a large amount of variance, depending on the sample. The expectation here, despite only seeking improvement, is also grounded in the fact that user generated content on social media is highly unreliable, can include sarcasm and be overall hard to process in terms of sentiment analysis. Nevertheless, the following hypothesis will be explored:

H2: Twitter sentiment score will increase the explanatory power of the Fama & French 3 and 5-factor model, on a portfolio level, for selected tech stocks

3. Data

3.1 Company specific data

Data regarding the stocks comprising the technology portfolio, namely *Apple, Google, Tesla, Microsoft, Amazon* is collected through *Yahoo Finance,* using a Python package called *yahoo-fin,* which ports the data from *Yahoo Finance*'s backend servers straight into Python. The data for the 5 companies is collected on a time frame of 2015 to 2020, to match the specific subset of the Twitter data collected for these companies. The data, since the analysis is done on a monthly timeframe, is resampled by keeping the last available datapoint for each company, for each month. From *Yahoo Finance,* the only relevant variable collected for the purpose of this research is the adjusted close price. In the context of applying the Fama and French 3 and 5 factor models, the adjusted close price is used for calculating monthly stock returns, which serve as the dependent variable in the models. The monthly returns variable is calculated using the following formula:

$$\text{Monthly Return}_t = \frac{\text{Adjusted Close}_t - \text{Adjusted Close}_{t-1}}{\text{Adjusted Close}_{t-1}}$$

Table 1: Descriptive statistics of returns for stocks comprising the technology portfolio

	AAPL	AMZN	TSLA	MSFT	GOOG
Mean	0.0976	0.1593	0.0913	0.1161	0.0860
Std	1.5646	1.8426	2.8312	1.4686	1.5122
Min	-9.9607	-7.8197	-13.9015	-9.2534	-7.6966
25%	-0.5857	-0.6551	-1.2852	-0.5401	-0.6029
50%	0.0893	0.1305	0.0454	0.0868	0.0629
75%	0.8918	0.9935	1.6247	0.8054	0.8257
Max	7.0422	14.1311	17.6692	10.4523	16.0524

Figure 1: Cumulative monthly returns between 2015 and 2020 for Microsoft, Tesla, Apple, Amazon and Google (Note the cumulative positive performance across the sample)

In figure 1 above, the cumulative returns for each of the stocks in the portfolio are graphed over the 5-year period being analyzed. From the graph, all the stocks selected exhibited positive returns, with all of them besides *Tesla* returning more than a 100% return. Clearly, *Amazon* outperformed the rest by a significant margin, with a cumulative return of slightly over 500%. Looking at table 1, where the monthly returns are presented in a descriptive statistics table, the statistics support the graph with *Amazon* having the highest mean return of 0.1593%. *Tesla's* returns have the highest standard deviation at 2.8312, a proxy for volatility, since its returns have the highest range, ranging from a low of –13.9015% to a high of 17.6692% in a given calendar month. The rest of the stocks exhibited similar levels of volatility, hovering around a standard deviation of 1.5. All of the companies in model, however, are analyzed on a portfolio level.

3.2 Fama & French 3 and 5 factor models

To analyze the stock returns using the Fama & French 3 and 5 factor models, data on the wider stock market must be incorporated as well. For the construction of the models, multiple variables

are considered to explain the variation in stock returns of the technology portfolio. The 3-factor model incorporates market risk, size and value as independent variables. The 5-factor model extends this by adding two more factors: profitability and investment. Specifics on how these variables are constructed will be discussed in subsequent sections of the paper. Nevertheless, these factors are used to construct regression models where the dependent variable is the excess return of the individual stocks in the technology portfolio. Since these parameters are market specific, and do not necessarily pertain to specific company data, a pre-existing library can be referenced for the data. The database including the Fama and French 3 and 5-factor model parameters references is compiled by the co-author of the models on his personal website *Kenneth R. French,* associated with Darthmouth University, which derives its information from CRSP financial database (French, 2023).

3.3 Twitter data

Twitter data for the 5 companies in question: *Apple, Google, Tesla, Microsoft, Amazon* were collected between 2015 and 2020 using a dataset from Kaggle, a popular data science platform which stores numerous datasets for the purposes of them being used in various machine learning applications, both in research and competition. In particular, this dataset was presented in a paper trying to establish a correlation between Twitter activity and trading volume (Doğan et al., 2020). The dataset includes 4,366,442 tweets with the ticker of each of the five companies referenced somewhere in the body of the tweet, on average providing more than 72,774 tweets per month from which sentiment can be derived. On Twitter, the way of specifying what you are talking about is called a *hashtag,* with individuals on the platform that typically reference the financial aspect of a company using the company stock's ticker symbol as a *hashtag.* The dataset is structured with multiple columns, including *tweet ID*, *writer*, *postdate*, *tweet body, ticker symbol,* as well as engagement metrics. This multiplicity of the variables provides a varied dataset able to be used in numerous different analyses, however, for the purposes of this research only the *tweet body* and the *ticker* will be used to extract sentiment and group it based on the company. In addition, all tweets included in the dataset either are or were at some point publicly available, and no personally identifiable information is used in the research, adhering to ethical guidelines of fair usage.

4. Methodology

4.1 Fama & French 3 and 5 factor models

The construction of the parameters for both the 3 and 5 factor models is done using a rigorous methodology, where data is collected both on the general stock market, as well as the US treasury bond market to compute the following parameters. The following two equations showcase the model parameters:

Model 1: Fama and French 3 factor model equation

$$R_{it} - R_{ft} = \alpha_i + \beta_1 \times (R_{Mkt,t} - R_{ft}) + \beta_2 \times SMB_t + \beta_3 \times HML_t + \epsilon_{it}$$

Model 2: Fama and French 5 factor model equation

$$R_{it} - R_{ft} = \alpha_i + \beta_1 \times (R_{Mkt,t} - R_{ft}) + \beta_2 \times SMB_t + \beta_3 \times HML_t + \beta_4 \times RMW_t + \beta_5 \times CMA_t + \epsilon_{it}$$

In model 1, representing the equation of the Fama & French 3-factor model, three independent variables are used to explain the variation in stock returns. The first is market risk, represented by the market return minus the risk-free rate. The market return is generally calculated using a value-weighted return of all stocks contained in the CRSP database that are listed on either NYSE, AMEX, or NASDAQ and that do not have missing data. The risk-free rate is the yield on a 1-month US Treasury bill, expressed as a monthly return. The difference between the market return and the risk-free rate thus gives the market risk premium. The size factor, represented by *SMB*, Small Minus Big, is constructed by taking the difference between the returns of small-cap and large-cap portfolios, which are sorted based on their market capitalizations, averaged out and then subtracted from each group. The value factor, known as *HML*, High Minus Low, is constructed by subtracting the difference between the returns of portfolios with high book-to-market ratios and those with low book-to-market ratios from the same dataset. These three variables serve as independent variables in an *ordinary least squares* (OLS) regression model where the dependent

variable is the excess return of the portfolio of stocks, represented by the difference between the stock return and the risk-free rate (French, 2023). The model thus by nature assumes a linear relationship between the dependent and independent variables, with an equal distribution of errors across the sampled data.

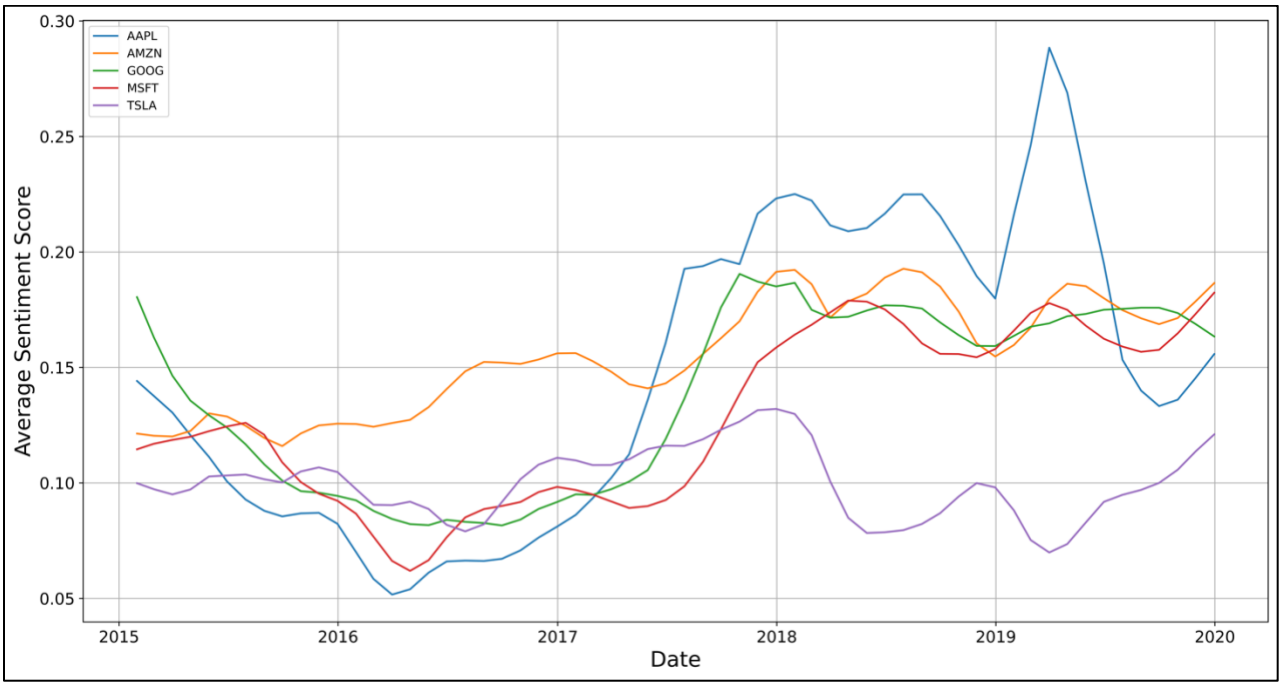
Similarly, model 2, representing the equation for the 5-factor model extends the 3-factor model by adding two more variables, profitability and investment. Profitability, represented as *RMW*, Robust Minus Weak, is calculated by taking the difference between the average returns of portfolios with robust profitability and weak profitability, measured using operating profitability. Investment, denoted as *CMA*, Conservative Minus Aggressive, is calculated by taking the difference between the average returns of portfolios with conservative and aggressive investment policies, measured using asset growth as a proxy for investment. These additional factors aim to capture variations in stock returns that are not explained by the original 3-factor model, with varying levels of success, as discussed in the theoretical framework.

4.2 VADER sentiment analysis

To put the twitter data to use when it comes to improving the predictive power of the two already well-established models above, a quantitative way of distilling text into data must be used. As described above, *VADER* (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool that is specifically engineered to analyze text from social media will be used to construct the sentiment scores of each tweet in the database. This technique, developed by a group of MIT data science researchers, is highly effective in analyzing short and informal text, as well as understanding both the polarity (positive / negative dimensions) of the sentiment expressed, but also its intensity (strength). A lexicon approach essentially uses a predefined list of words and phrases where each data point is assigned a sentiment score manually. In this case, the scores range from extreme negative (-4) to extreme positive (+4), For example, the word “brilliance” is assigned a score of 2.9, while “criminal” is assigned a –2.4, which are then used as references when analyzing new text, where words are very likely to overlap. Following this, the scores are normalized to a scale between –1 and 1, where a higher score suggests a more positive

sentiment, with 0 being neutral. In addition, the lexicon also includes slang, emoticons, and other social media-specific terms, making it well-suited for analyzing user generated content from social media. This sentiment analysis tool is deployed using a package in Python, assigning a unique sentiment score to each tweet separately, after which the scores are grouped by company and simply averaged out on a per month basis.

Figure 2: Moving Average of Compound Sentiment Scores across all 5 portfolio companies



In figure 2 above, the moving average of the compound sentiment scores, per month and grouped for each firm being analyzed are graphed. The sentiment scores range between approximately 0.05 and 0.3, indicating a generally positive sentiment across all companies, across the dataset. Generally speaking, this makes sense, given that over the course of the timeframe that is being analyzed here, all the companies exhibited positive compound returns. Interestingly, however, the sentiment never approached 0 (neutral), or a negative score, meaning that on average, even when the stock was performing poorly, the social media sentiment remained somewhat positive, on average.

4.3 Expanded Models

Model 3: Fama and French 3 factor model equation with a sentiment score

$$R_{it} - R_{ft} = \alpha_i + \beta_1 \times (R_{Mkt,t} - R_{ft}) + \beta_2 \times SMB_t + \beta_3 \times HML_t + \beta_4 \times SEN_t + \epsilon_{it}$$

Model 4: Fama and French 5 factor model equation with a sentiment score

$$R_{it} - R_{ft} = \alpha_i + \beta_1 \times (R_{Mkt,t} - R_{ft}) + \beta_2 \times SMB_t + \beta_3 \times HML_t + \beta_4 \times RMW_t + \beta_5 \times CMA_t + \beta_6 \times SEN_t + \epsilon_{it}$$

Using the compound monthly sentiment scores for each company derived above, a sentiment score variable can be added to both Fama & French models as is, given that it is a continuous variable with a score between 0 and 1. The addition of the variable creates two new OLS models, which will combine the preexisting data with the sentiment score. These models will be compared to the normal Fama & French models using an F-test, as well as a comparison of their respective R and R^2 values, to see whether the sentiment variable improves the explanatory power and the model fit.

5. Results

5.1 Fama & French 3 and 5 factor model results

In table 2, the OLS results for both the Fama & French 3-factor model (above as model 1) and the Fama & French 5-factor model (above as model 2) are presented. For both models, the left-hand side of the equation for both models is the variable *excess returns*, defined as the difference between portfolio returns and market returns. Running the OLS regressions for all 5 companies, across 60 months (5 years) of data, using the previously explained returns yields the following results.

Table 2: Regression outputs of OLS models for Fama & French 3 and 5-factor models

	Fama & French 3-factor model (1)	Fama & French 5-factor model (2)
Constant	0.0116**	0.0121***
	(0.005)	(0.004)
Mkt - RF	1.1048***	0.9661***
	(0.131)	(0.133)
SMB	-0.4316**	-0.4514 **
	(0.197)	(0.203)
HML	-0.3522*	0.0545
	(0.178)	(0.217)
RMW		-0.1259
		(0.337)
CMA		-1.0400***
		(0.363)
Observations	60	60
R^2	0.581	0.640
Adjusted R^2	0.559	0.607

Notes: Write some notes here. All the coefficient values are reported with the significance level represented by asterisks, as such: *significant at 10% ($p < 0.1$), **significant at 5% ($p < 0.05$), ***significant at 1% ($p < 0.01$)

Looking at the R^2 values, for both models, the expanded 5-factor model has a slightly better fit to the data. The R^2 value for the 3-factor model is 0.581, while the 5-factor model has an R^2 value of 0.640. Looking at the adjusted R^2 values, factoring in the number of variables in the model which can bias the R^2 value towards 1, also indicate that the 5-factor model explains more of the variance (adjusted $R^2 = 0.607$) as compared to the 3-factor model (adjusted $R^2 = 0.559$). This is evidence in support of the alternative hypothesis (H1), stating that the 5-factor models explain a high variation of excess returns.

Looking at the coefficients of the models, the constant term (α) is significant in both, with the 5-factor model exhibiting a slightly higher one of 0.0121, as compared to 0.0116. With Fama & French models, the α represents the excess return after accounting for market risk (in the models presented as the *Mkt – RF* variable), as well as all other factors. Given that the α is positive in both models, it suggests that the portfolio outperformed the expected returns, given the systematic risk. The coefficients for the market risk are also positive in both models, more so for the 3-factor model with 1.1048, than the 5-factor model with 0.9661. With the coefficient being close to 1, it suggests that the portfolio of these stocks moves in line with the market, since for every 1% point increase in market risk, the excess returns go up by 1% point as well. Seeing that the 5-factor model has a lower market risk coefficient, it suggests that additional factors capture the variances that are otherwise attributed to market risk.

The coefficients of SMB are negative in both and statistically significant, with values of -0.4316 for the 3-factor and -0.4514 for the 5-factor model. This suggests that technology stocks perform worse in times when stocks with smaller market capitalizations perform better, making sense since all of them have high market capitalizations. HML, being -0.3522 and significant only in the 3-factor model at the 10% significance level, would imply that the portfolio tends to perform worse when value stocks outperform growth stocks, however, due to the insignificance cannot be regarded as such. The remaining two factors, RMW and CMA, pertain only to the 5-factor model. The RMW coefficient is negative and not statistically significant at any level, implying that profitability is not a strong predictor of excess returns for the constructed technology portfolio. The CMA on the other hand, being highly significant and negative at –1.04, suggests that the portfolio performs better when stocks with aggressive investment strategies outperform stocks with conservative investment strategies, likely meaning that the firms in the portfolio themselves are investing aggressively.

The evidence, based on the interpretation of the statistical models above, allows for a conclusion to be made regarding the first hypothesis, stating that the 5-factor model performs better when it comes to explaining the variance in the excess returns of the selected technology stocks. The results are robust after inspecting metrics such as R^2, adjusted R^2 and the additional significant factor in

the 5-factor model, the CMA factor, providing a better understanding of the drivers of excess returns in the technology sector. The 5-factor Fama & French model thus explains a higher portion of the variance in excess returns for the selected technology stocks, as compared to the 3-factor model, supporting the alternative hypothesis.

Table 3: Regression outputs of OLS models for Fama & French 3 and 5-factor models with a sentiment variable

	Fama & French 3-factor + sentiment (1)	Fama & French 5-factor + sentiment (2)
Constant	0.0185 (0.019)	0.0153 (0.026)
Mkt - RF	1.1084 *** (0.133)	0.9663*** (0.133)
SMB	-0.4357 ** (0.199)	-0.4528 ** (0.205)
HML	-0.3655 * (0.183)	0.0519 (0.220)
RMW		-0.1234 (0.341)
CMA		-1.0514*** (0.363)
SEN	-0.0525 (0.141)	-0.0141 (0.112)
Observations	60	60
R^2	0.582	0.640
Adjusted R^2	0.552	0.599

Notes: Write some notes here. All the coefficient values are reported with the significance level represented by asterisks, as such: *significant at 10% ($p < 0.1$), **significant at 5% ($p < 0.05$), ***significant at 1% ($p < 0.01$)

5.2 Fama & French 3 and 5 factor model results with sentiment variable

Table 3 presents the OLS results of the expanded Fama & French models, with the added sentiment variable, *SEN*, capturing the social media beliefs regarding the technology stocks in the underlying portfolio. Here, across the same time span of 5 years, the additional variable is added to the models as normalized variable, with values between 0 and 1. By adding this variable, the models are now attempting to capture the effects of investor sentiment on excess stock returns. The R^2 and adjusted R^2 values remain very similar to the values reported in table 2, with the adjusted R^2 decreasing by 0.007 for the 3-factor model and by 0.008 for the 5-factor model. The implication here is that the inclusion of the sentiment variables does not help to explain any additional variation of the excess returns, over what is already explained by the constructed variables in the models.

While the coefficients for the sentiment variable in both models are negative, which would imply that a more negative underlying sentiment regarding the stocks would lead to higher excess returns, the variable is insignificant in both models. The results are insignificant at the 5% significance level, suggesting that the cumulative sentiment of the collected Twitter data is not a statistically significant predictor of excess returns for the chosen set of technology stocks. In addition, the inclusion of this variable does not substantially change the coefficients or their significance levels of any of the other variables, suggesting that the effects of those variables on returns remain robust after the addition of investor sentiment in the model. In other words, for the sample analyzed here over the given period, it appears that fundamental factors influence returns significantly more than investor sentiment.

6. Discussion

6.1 Fama & French 3 vs 5 factor model discussion

As seen in previous research, the general scientific consensus surrounding the debate whether the two additional factors (*profitability* and *investment*) contribute to the expanded model's ability to capture the variance in excess returns has been mixed. While the original paper (Fama & French, 2015) that coined the model claiming an improvement over their previous 3-factor model, there

have been numerous studies arguing that the 5-factor model increases complexity at the expense of a non-guaranteed improvement. For instance, Cakici (2015) found that the two additional factors do not add any explanatory power in stock portfolios constructed from predominantly Asian stocks. Here, however, for a select group of US technology-oriented stocks, the 5-factor does seem to perform better than the 3-factor model, adding to the body of research in favor of the expanded model. It should be noted, however, that even in the original paper, it has been suggested that the 5-factor model tends to perform better in a well-diversified portfolio, comprised of stocks with different characteristics. In addition, Artman et al. (2012) found that the 3-factor model (hence, the 5-factor by nature as well) has been found to poorly explain the returns of average performing stocks in certain markets. This suggests that a part of the reason why the models work well in this study is because they have all had above average returns over the period, represented by both models having a statistically significant, positive alpha. While the statistical results are in favor of the hypothesis that the 5-factor model performs better than the 3-factor model, the results should be extrapolated cautiously due to the portfolio analyzed here being a very small, unrepresentative subsection of the US stock market.

6.2 Sentiment variable insignificance & Data limitations

With the implementation of the sentiment variable in the Fama & French's models, the expectation was conservative, but based on previous research optimistic. Given that it is widely accepted that numerous behavioral biases are present in the current financial market, such as Li et al. finding the presence of *hearding*, especially in individual investors, the notion that psychological traits of investors can affect the stock market are not too farfetched (2023). In addition, previous research has found links between activity on social media and the movement of the financial market (Ranco et al., 2015). Here the idea of using Twitter data to try proxy sentiment and try to explain a high portion of the variation of excess returns in Fama & French models does not support similar research. While Bollen, Mao & Zheng have found that twitter sentiment can accurately predict the closing value of ETFs, a similar approach for the 5 technology stocks analyzed here does not yield the same results (2011). This suggests that using Twitter can be useful in analyzing wide market movements, but falls short in more specific company analysis, however, this is highly contextual, and more research is needed.

When it comes to the method of mediating Twitter data and the investor sentiment through there are also numerous shortcomings in the data that can lead to poor or insignificant results. Firstly, using Twitter data cannot by itself capture the entire investor sentiment surrounding the securities being analyzed, since numerous institutional investors and older individual investors do not necessarily participate in social media, making it difficult to accurately extract a representative sentiment. Twitter and other social media platforms also suffer from manipulation from bots that introduce random noise in the data, as well as from individuals whose actual view and beliefs do not align with their activity, due to sarcasm or deliberate misinformation. Also noteworthy is that it is virtually impossible to know whether the individual users tweeting surrounding the assets are investing in them as well, making it impossible for the data to accurately represent *investor* sentiment. In addition, a possible explanation why previous research has found success in shorter time frames of analysis is due to the nature of fast moving, in the moment discussion that happen on Twitter, making it not as good of a predictor for long-term market movements.

6.3 Suggestions

For future research, a deeper exploration in two areas is suggested. First one exploring whether the Fama & French 5-factor model outperform the 3-factor model across different industries, as well as different asset market. While the research presented here provides evidence that it does, it is limited to only 5 US companies in the technology section, making it unrepresentative for firm conclusions to be drawn. The second suggestion has to do with investor sentiment, since previous research has been able to establish a link between sentiment and market movements, there is potential for future studies to further the connection. Here, research has seen positive findings both regarding wider market movements, such as that of ETFs in Bollen, Mao & Zheng (2011), as well as specific more specific, one-off events. Using Reddit data for a single stock: GameStop (GME), Wang & Luo (2021) were able to establish a partial correlation between sentiment and the stock price movement using VADER, on a daily time frame. Future research should also address the direction of the causality in instances where it does prove to be significant, since positive price movements could lead to a more positive sentiment, and not necessarily the other way around.

7. Conclusion

The main theme of this study is to explore the impact of Twitter derived investor sentiment on the predictive power of the Fama & French 3 and 5-factor models. The sample consisted of five predominantly technology stocks: *Tesla*, *Apple*, *Google*, *Microsoft*, and *Amazon* with data collected between 2015 and 2020, as well as a Twitter text corpus consisting of over four million tweets regarding these companies. Using this, two hypotheses were explored, firstly that the Fama & French 5-factor model would explain a higher portion of the variance when compared to its 3-factor counterpart, and second, that adding a sentiment variable based on the twitter data would increase the predictive power of both models. Hence, the main question attempted to be answered here was whether investor sentiment can help explain the variance in excess returns.

To examine the performance of the two models, in addition to seeing whether the sentiment variable improved efficiency of the models, *Ordinary Least Squares* (OLS) regression was used. This is because the CAPM model, the underlying model of the Fama & French literature proposes a linear relationship between risk and return, hence a linear regression model is used. The study was conducted over a 5 year period, from 2015 to 2020, where it was concluded that the 5-factor model explains a slight, but a statistically significant higher portion of the variance in the excess returns, when compared to the 3-factor model. Using more than 4 million tweets regarding the 5 companies in the portfolio to form the sentiment variable, in addition to the model parameters, was shown to be insignificant in both models. The study offers some industry specific insights into asset pricing theory, as well as doubts about the use of Twitter data in finance. The study shows that in the technology sector, the expanded Fama & French 5-factor tends to outperform the 3-factor model, adding to the body of research in favor of the 5-factor model, at least in the technology industry. In addition, the study highlights the difficulty and the potential unreliability of using Twitter data in hopes of improving asset pricing models. While investors' expectations and beliefs are likely to play a role in asset pricing and returns, further studies must be conducted to better proxy it, in hopes of gaining explanatory power.

References

Artmann, S., Finter, P., & Kempf, A. (2012). Determinants of expected stock returns: large sample evidence from the German market. Journal of Business Finance & Accounting, 39(5-6), 758-784.

Bartholdy, J., & Peare, P. (2005). Estimation of expected return: CAPM vs. Fama and French. International Review of Financial Analysis, 14(4), 407-427. https://doi.org/10.1016/j.irfa.2004.10.009

Belen Blanco. (2012). The use of CAPM and Fama and French three factor model: portfolios selection. Public and Municipal Finance, 1(2).

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

Cakici, N. (2015). The Five-Factor Fama-French Model: International Evidence. Available at SSRN: https://ssrn.com/abstract=2601662

Corea, F. (2016). Can Twitter proxy the investors' sentiment? The case for the technology sector. Big Data Research, 4, 70–74. https://doi.org/10.1016/j.bdr.2016.05.001

Costola, M., Hinz, O., Nofer, M., & Pelizzon, L. (2023). Machine learning sentiment analysis, COVID-19 news and stock market reactions. Research in International Business and Finance, 64, 101881. https://doi.org/10.1016/j.ribaf.2023.101881

Doğan, M., Metin, Ö., Tek, E., Yumuşak, S., & Öztoprak, K. (2020, December). Speculator and influencer evaluation in stock market by using social media. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 4559-4566). IEEE.

Elbannan, M. (2014). The capital asset pricing model: An overview of the theory. International Journal of Economics and Finance, 7, 216. https://doi.org/10.5539/ijef.v7n1p216

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3–56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116(1), 1-22.

French, K. (2023). Data Library. Tuck School of Business at Dartmouth. Retrieved from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Galagedera, D. U. A. (2007). A review of capital asset pricing models. Managerial Finance, 33(10), 821-832. https://doi.org/10.1108/03074350710779269

Gandhi, P., et al. (2020, September 14). Which industries are the most digital? Harvard Business Review. https://hbr.org/2016/04/a-chart-that-shows-which-industries-are-the-most-digital-and-why

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance, 48(5), 1779-1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

Griffin, J. M. (2002). Are the Fama and French factors global or country-specific? The Review of Financial Studies, 15(3), 783-803.

Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019). Sentiment analysis with NLP on Twitter data. In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). https://doi.org/10.1109/IC4ME247184.2019.9036670

Houlihan, P., & Creamer, G. G. (2017). Can sentiment analysis and options volume anticipate future returns? Computational Economics, 50, 669–685. https://doi.org/10.1007/s10614-017-9694-4

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In E. Adar et al. (Eds.), ICWSM. The AAAI Press.

Jiao, W., & Lilti, J. J. (2017). Whether profitability and investment factors have additional explanatory power comparing with Fama-French Three-Factor Model: empirical evidence on Chinese A-share stock market. China Finance and Economic Review, 5(7). https://doi.org/10.1186/s40589-017-0051-5

Leković, M. (2020). Cognitive biases as an integral part of behavioral finance. Economic Themes, 58(1), 75-96.

Li, T., Chen, H., Liu, W., Yu, G., & Yu, Y. (2023). Understanding the role of social media sentiment in identifying irrational herding behavior in the stock market. International Review of Economics & Finance, 87, 163-179.

Lo, A. W. (2007). Efficient markets hypothesis. The New Palgrave: A Dictionary of Economics. https://ssrn.com/abstract=991509

Lucey, B., Xie, Y., & Yarovaya, L. (2023). "I just like the stock": The role of Reddit sentiment in the GameStop share rally. Financial Review, 58(1), 19-37. https://doi.org/10.1111/fire.12328

Maiti, M. (2020). A critical review on evolution of risk factors and factor models. Journal of Economic Surveys, 34(1), 175-184.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60).

O'Connor, A. J. (2013). The power of popularity: An empirical study of the relationship between social media fan counts and brand company stock prices. Social Science Computer Review, 31(2), 229-235.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. PloS one, 10(9), e0138441.

Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis.

Shiva, A., & Singh, M. (2020). Stock hunting or blue chip investments? Investors' preferences for stocks in virtual geographies of social networks. Qualitative Research in Financial Markets, 12(1), 1-23. https://doi.org/10.1108/QRFM-11-2018-0120

Smith, S., & O'Hare, A. (2022). Comparing traditional news and social media with stock price movements; which comes first, the news or the price change? Journal of Big Data, 9(1), 1-16. https://doi.org/10.1186/s40537-022-00591-6

Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012, December). An experiment in integrating sentiment features for tech stock prediction in twitter. In Proceedings of the workshop on information extraction and entity analytics on social media data (pp. 23-38).

Wang, C., & Luo, B. (2021). Predicting $ gme stock price movement using sentiment from reddit r/wallstreetbets. In Proceedings of the Third Workshop on Financial Technology and Natural Language Processing (pp. 22-30).

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Zhang, C. (2008). Defining, modeling, and measuring investor sentiment. University of California, Berkeley, Department of Economics.