FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

PREDICTING MEME STOCK PRICES USING RECURRENT NEURAL NETWORKS AND SENTIMENT ANALYSIS ON WALLSTREETBETS

Aantal woorden: 18.355

Lars Cuvelier
Stamnummer: 01503525

Promotor: Prof. Dr. Els Clarysse

Masterproef voorgedragen tot het bekomen van de graad van:

Master in de handelswetenschappen: management en informatica

Academiejaar: 2021-2022

UNIVERSITEIT GENT

Abstract

Meme stock valuations rose sky-high in early 2021. At the same time, the Reddit forum of WallStreetBets was in ecstasy. Their user base of retail investors claimed responsibility for the meme stock bull run. This study aims to reliably predict five meme stock share prices by analyzing sentiment on WallStreetBets using multiple linear regression (MLR) and Gated Recurrent Unit (GRU) neural networks. We gather our data by web scraping directly off WallStreetBets. Our dataset consists of 1,456,167 user posts from 2020 up to 2022. A support vector machine sentiment analysis model categorizes sentiment of the posts. Stock market data from the same period merged with sentiment analyzed WallStreetBets data are the independent variables in our two prediction methods. The closing share price of the meme stocks is the independent variable. Three out of the five meme stocks MLR models are strong and reliable prediction models with R-squared values above 0.843. The best performing MLR model predicts Tesla's share price with an R-squared value of 0.944. The other prediction model is the three-layered GRU model, which is a type of recurrent neural network. All the five GRU models are unreliable in predicting the next day's share price. The results suggest that predicting the volatile meme stocks is challenging. The daily average sentiment of the WallStreetBets forum is only intermittently a significant contributing variable to the MLR prediction models. Additionally, the meme stocks' volatile valuations from 2020 to 2022 might not be a suitable period for training the prediction models. Literature on WallStreetBets as an influential group of retail investors and the meme stocks phenomenon is lacking. Further research is welcomed on these topics, as well as the use of GRU as a stock prediction model.

Acknowledgments

I would like to thank my promotor Els Clarysse for her patience, guidance and feedback. I am thankful for the opportunity to research this topic and to be able to fully choose the approach I desire.

I would also like to express my gratitude to my parents for their continued support and healthy pressure throughout my education and thesis. Their advice and feedback have been extremely helpful.

Table of Contents

List of Abbreviations

API	Application Programming Interface
BB	BlackBerry
BOW	Bag-of-Words
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma Separated Values
DCF	Discounted Cash Flow
EBITDA	Earnings Before Interest, Taxes, Depreciation, and Amortization
EPS	Earnings Per Share
EV	Enterprise Value
FCFE	Free Cash Flow to Equity
FCFF	Free Cash Flow to the Firm
FN	False Negatives
FP	False Positives
GRU	Gated Recurrent Unit
LSTM	Long short-term memory
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Squared Error
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
NN	Neural Networks
P/E	Price to Earnings
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
Sklearn	Scikit-learn
SVM	Support Vector Machine
TFIDF	Term Frequency–Inverse Document Frequency
TN	True Negatives
TP	True Positives
UTF-8	8-bit Unicode Transformation Format
WACC	Weighted Average Cost of Capital
WSB	WallStreetBets

List of Figures

List of Tables

1. Introduction

GME to the moon! On the 27th of January 2021, the share price of GameStop (GME) closed at $347. Just one month before that, the same stock was trading below $25. The amazing run of GameStop on the stock market attracted a lot of attention from investors, media and financial institutions. Many seemed puzzled that GameStop, a company who had suffered gravely during the COVID-19 pandemic, could attract so many investors leading to an unprecedented all-time high share price. However, the users from the social media platform Reddit were far from surprised. A forum dedicated to risky stock market investing called WallStreetBets had experienced the exponential growth of GameStop's share price firsthand. WallStreetBets' users enthusiastically shared their ever-increasing returns on GameStop every day. The stock prices eventually rose so ridiculously high that it was dubbed as a meme stock, a meme being an internet joke. In fact, GameStop's share price reached levels so elevated that financial institutions risked shorting the stock, resulting in an all-out conflict between the online forum of Reddit and financial institutions. The meme stock conflict popularized WallStreetBets in the public and within the social media platform of Reddit itself.

In this research, we will try to predict meme stocks' share prices with two different models and data from the WallStreetBets forum. First, we will analyze the forum's daily sentiment of five popular meme stocks, including GameStop. The sentiment analysis is realized by machine learning models that categorize WallStreetBets' posts in positive and negative sentiment. The daily average sentiment per meme stock serves as one of the variables to two different machine learning models that try to predict meme stock share prices. These two models are multiple linear regression and gated recurrent unit, a type of recurrent neural networks. Overall, our concluding research question is:

Can multiple linear regression or gated recurrent unit serve as a reliable model to predict meme stocks share prices by analyzing sentiment on Reddit's WallStreetBets?

In this paper, we start by reviewing the macroeconomic context of COVID-19 pandemic in the literature review, followed by an introduction to Reddit's WallStreetBets and the meme stock phenomenon. In addition, methods for stock price valuation are established, as well as every machine learning method and model that is used in this research. Next, we discuss the methodology that serves as the structure for the data mining process. The results are discussed separately and are followed by the discussion, where the results are analyzed and contextualized. Finally, the findings and process of this research paper are wrapped up in the conclusion.

2. Literature review

2.1 COVID-19 pandemic macroeconomic indicators & uncertainty

The economic fallout of the COVID-19 pandemic was unprecedented in scale and speed. Altig et al. (2020) considers numerous economic indicators and measures of uncertainty during the COVID-19 pandemic in the US and the UK. The economic indicators presented by the authors aid to comprehend the immensity of the crisis. The stunning surge of unemployment rates is one of the most prominent effects of the COVID-19 pandemic. The new claims for unemployment benefits in the US had seen an increase from a weekly average of 209,250 in mid-February 2020 to a staggering 5,301,250 in mid-April. At the same time, the US unemployment rate jumped from a record low of 3.5% to an 80-year high of 14.7%. Another economic indicator that further illustrates the incredible consequences of the COVID-19 pandemic is the GDP output. The global economic contraction in the second quarter of 2020 were the first of its kind in the modern era. The United States' GDP plunged 11.2% from 2019 Q4 to 2020 Q2, while in April 2020, the UK experienced the greatest ever monthly drop in GDP with an astounding 20.4%. These concerning key macroeconomic indicators all demonstrate the severity of the crisis.

Altig et al. (2020) further argue that the economic response to the COVID-19 pandemic is unprecedented, primarily due to the colossal scale and abruptness of the economic contraction. As a result, the measured levels of uncertainty soared. An influential uncertainty measure is the stock market volatility. The volatility index (VIX) is the most popular measure of the stock market's expectation of volatility in the near-term future (Saha et al., 2018). The 1-month VIX index gauges the implied 30-day volatility of the market, calculated from options on the S&P 500. The index tends to rise when the stock market performance is expected to weaken. In early January 2020, the 1-month VIX index maintained steady values within the 10 to 15 range. On the 16th of March 2020, the daily VIX rose to a peak value of 82.7. In comparison, the highest value of the VIX index during the Global Financial Crisis of 2008 was 80.9.

Measure	**Value in January 2020**	**% Jump Jan 2020 to Peak**	**Date of COVID-19 peak value**
VIX 1-month implied volatility[1]	13.3	497	March 16
News Economic Policy Uncertainty, US[2]	110.1	683	May 26
Twitter Economic Uncertainty, US[3]	139.8	594	April 22-28

Table 1: Different measures of uncertainty COVID-19 pandemic

[1] www.cboe.com/vix
[2] www.policyuncertainty.com
[3] Baker et. al., (2020)

A second uncertainty measure selected by Altig et al. (2020) is a newspaper-based uncertainty measure. Economic Policy Uncertainty (EPU) index introduced by Baker et al. (2016) tracks the frequency of the words "economics", "policy" and "uncertainty" in approximately 2000 US newspapers. The EPU index uses the average uncertainty measured from 1985 through 2010 as a normalized value of 100. In January 2020, the US EPU index maintained an average daily uncertainty value close to 100. Later, during the first months of the COVID-19 pandemic, the monthly US EPU soared to the highest values to date, reaching a peak value of 752 on the 26th of May 2020. The EPU index is a valuable measure to gauge journalists' perceived uncertainty and economic outlook. Another similar behaving index is the Twitter-based Economic Uncertainty (TEU) constructed by Altig et al. (2020). This index tracks tweets that mention the words and variants of "economic" and "uncertainty" from 2010 and onwards. Table 1 demonstrates the divergence between the base and peak values of the indices and the peak value date. Both the EPU and TEU indices peak considerably later than the VIX index, indicating a difference in perceived uncertainty levels by stock market participants compared to journalist and Twitter users. We will discuss the effects of the COVID-19 pandemic on the stock market next.

2.1.1 Stock market response to the COVID-19 pandemic

As the many alarm bells of the macroeconomic indicators and uncertainty measures went off, the stock market reacted heavily. The effect of the COVID-19 pandemic on the stock market has been unlike any previous infectious disease outbreak (Baker et al., 2020). The COVID-19 pandemic shares many similarities with the Spanish Flu pandemic. However, in terms of excess mortality rates, the Spanish Flu pandemic was at least ten times deadlier than COVID-19. Nonetheless, the Spanish Flu did not result in any major effects on the stock market. Throughout the 1918-1919 pandemic, not a single movement of more than 2.5% on the US stock market was recorded that was attributed to the economic fallout of the pandemic. The discrepancy between both pandemics and its impact on the stock market could not be greater. The economic fallout of the COVID-19 pandemic alone triggered more than 12 major daily moves on the US stock market (Baker et al., 2020). From early February to the 23rd of March 2020, the S&P 500 temporarily halted trading three times when the index dropped more than 7% and had lost more than a third of its valuation.

In the first wave of the COVD-19 pandemic in March 2020, extremely harsh measures were taken by governments around the world to prevent the spreading of the coronavirus. Baker et al. (2020) suggests that the government restrictions on commercial activities in combination with social distancing in a modern service-driven economy had a great effect on the stock market. Starting in March 2020, most countries issued heavy restrictions on movement and commercial activities. The aggressive COVID-19 measures in the US caused at least 70% of its residents to receive stay at home orders and experience

mandatory closing of essential businesses. Weakly flight frequency dropped by 75% during the second quarter of 2020 whilst transporting less passengers per flight. The restrictions issued by governments combined with social distancing drastically reduced overall economic activity. Additionally, the social distancing measures intensified the economic fallout due to the increasing service-driven nature of modern economies. Baker et al. (2020) concludes that the combination of government restrictions on free movement and widespread mandatory closures of businesses in the ever-growing service oriented modern economy are the likely reasons why the economy and stock market has experienced such an enormous contraction.

2.1.2 Rebound and investor behavior to the stock market crash

While the stock market losses in response to the COVID-19 pandemic were enormous, the S&P 500 has taken only 126 days to fully recover from its lowest valuation during the COVID-19 pandemic, making it by far the fastest recovery in history. There are many possible causes to explain the speedy recuperation. Giglio et al. (2021) analyzed the change in investors' expectations and stock returns during the COVID-19 crash and its recovery. In the wake of the stock market crash, investor sentiment for the short-term stock market performance turned pessimistic. However, the expectations on the long term (10-year) largely remained constant or improved. Pagano et al. (2021) also reviewed investor behavior during the recovery of the COVID-19 stock market crash. The authors pinpoint the government mandated stay-at-home orders as one of the reasons why large amounts of equity made its way to trading platforms. While stocks were plummeting, Robinhood, a large zero cost trading service in the US, saw its average trading volume triple. These new investors appeared to be very responsive to news while also exhibiting contrarian behavior such as investing in airlines and cruise line stocks, which suffered gravely during the pandemic. Government mandated stay-at-home orders combined with moderately positive investor expectations in the long run and an influx of new investors are one of many reasons why the stock market rebounded in an unprecedented manner.

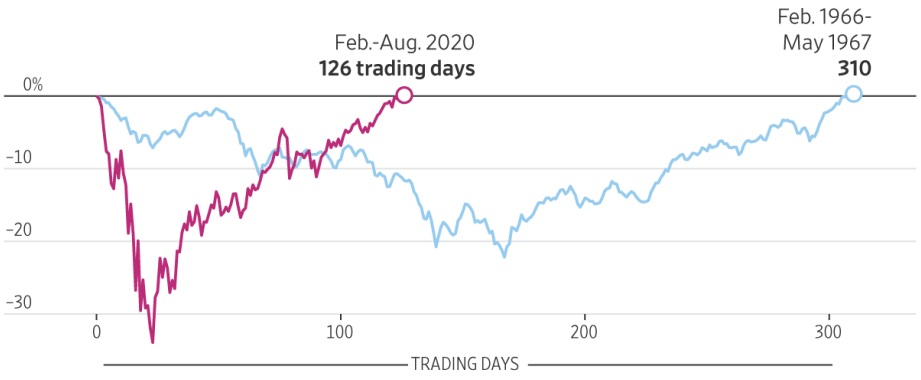


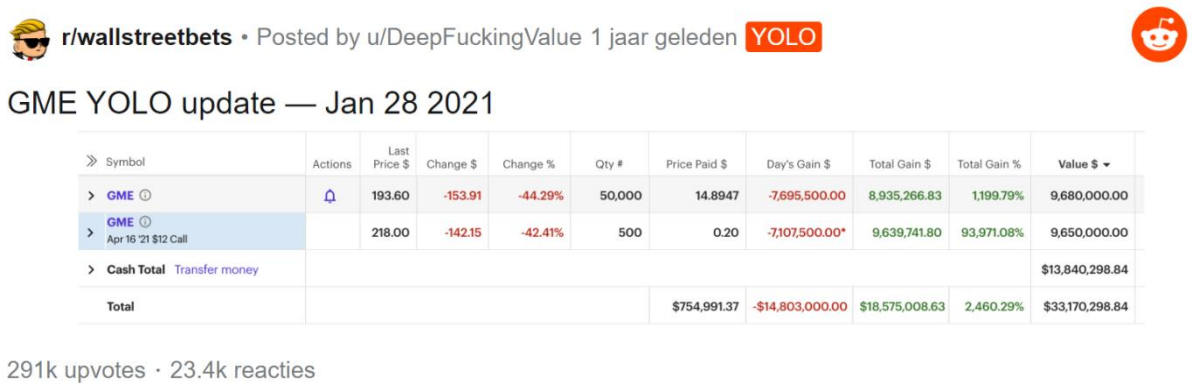
Figure 1: S&P 500 fastest recovery from bear market (Banerji, 2020)

2.1.3 New retail investors and investing apps

In the last few years, large numbers of retail investors have begun trading on the stock market due to the ease of use and the low fees of investing apps (Chaudhry & Kulkarni, 2021). Retail investors are defined as those who invest their own money and have no professional training in investing. Investing apps such as Robinhood, Public and Webull offer those retail investors free or low-cost services to trade stocks, options, and securities without requiring any investing training or minimum deposit. The investing apps claim that their aim is to democratize investing so that anyone can have access to stock market trading. The danger of lacking a minimal required knowledge of investing is that unexperienced retail investors may trade irrationally or emotionally. In addition, the design choices of the investing apps and the trading of stocks and options without fees might even encourage irrational behavior such as high frequency trading (Chaudhry & Kulkarni, 2021). As an example of design choice, Robinhood attracts new users with a 'Golden Ticket' lottery, whereby a random stock is deposited in the new user's portfolio as a reward for signing up. Such a deliberate design choice closely resembles gambling and appears to be effective in attracting new retail investors. New users that flock to investing apps largely have similar profiles. Pagano et al. (2021) labels these new retail investors post COVID-19 contrarian in behavior, accepting more risk than average investors. In some cases, the senseless risk that several retail investors had taken paid off enormously. Some outrageous returns were posted on the social media platform of Reddit and became very popular. On Reddit, there is a community of risky retail investors that applauds irrational investing decisions. That community is called WallStreetBets.

2.1.4 r/WallStreetBets and meme culture

WallStreetBets (WSB) is a subreddit on the popular social media platform Reddit. Users of Reddit can join different communities or subreddits according to their own interests. Users can post, comment or vote on Reddit. There are more than 100,000 communities and 50 million daily active users (Reddit.inc, 2022). Some popular subreddits indicated by the preceding 'r/' are r/funny, r/games or r/soccer. The subreddit of WallStreetBets has a following of 11 million members and describe themselves as a community 'for making money and being amused while doing it. Or, realistically, a place to come and up vote memes when your portfolio is down' (WallStreetBets, 2012). The humorous and nonchalant description is an appropriate characterization of the subreddit's user base. The most up voted and thus popular content posted to the subreddit are absurd returns or losses on stocks, clearly the result of high-risk irrational investment. Other popular content are memes of certain stocks or events. An internet meme is an idea or behavior, usually a humorous one, spread via social media (The Economist, 2021a).



Symbol	Actions	Last Price $	Change $	Change %	Qty #	Price Paid $	Day's Gain $	Total Gain $	Total Gain %	Value $
GME		193.60	-153.91	-44.29%	50,000	14.8947	-7,695,500.00	8,935,266.83	1,199.79%	9,680,000.00
GME Apr 16 '21 $12 Call		218.00	-142.15	-42.41%	500	0.20	-7,107,500.00*	9,639,741.80	93,971.08%	9,650,000.00
Cash Total Transfer money										$13,840,298.84
Total						$754,991.37	-$14,803,000.00	$18,575,008.63	2,460.29%	$33,170,298.84

Figure 2: Popular post of Keith Gill on r/ WallStreetBets. (Gill, 2021)

One of the most popular posts on r/ WallStreetBets is a January 2021 monthly update of the subreddits most celebrated member Keith Gill with the ridiculous Reddit username 'DeepFuckingValue'. In September 2019, Gill believed that the stock value of GameStop (GME) was severely undervalued (Prentice, 2021). He subsequently bought 50,000 shares and 500 call options at a value of $750,000, essentially betting on an increase in stock prices in the near to long term. By January 2021 the stock price had increased from $16 to $325. Gill's shares increased more than 2,000% in value, while his call options had a staggering return of 155,641%. In total, Gill's investment at their peak were valued at $46 million. His post on the subreddit was up voted by more than 290,000 Reddit-users. As a reference, the most liked post in the history of Reddit has 450,000 up votes in May 2022. That post is a video of a Reddit user that personally bought a New York Times Square Billboard on the 30th of January 2020 that said GME go BRRRR. The 'brrrr' is a meme that started out as a joke that makes fun of the US Federal Reserve. The meme's original format was: 'money printer go brrrr' and his its own website 'brrr.money' hosted by the satirical Institute for Memetic Research & Development. Its purpose was to phonetically imitate the sound of printing money. In essence, WallStreetBets Reddit as a community was ridiculing the amount of quantitative easing that was done during the COVID-19 pandemic.

2.1.5 GameStop and short squeeze

GameStop is one of the largest American based games retailers. They mainly sell video games, gaming consoles and collectibles. The company operates a brick-and-mortar strategy with 4800 stores worldwide. Over the last few years, market share dropped steadily to online games retailers (The Economist, 2021b). Then in 2020, the COVD-19 pandemic forced GameStop to temporarily close its stores. Revenue halved during the pandemic when compared to 2015-levels, and its share price continued to plummet. In 5 years, GameStop's share price fell from $35 to $2.8 per share. Investing in GameStop seemed like a terrible idea. However, the share price steadily rose from its trough in mid-2020 and even became parabolic near the end of 2020.

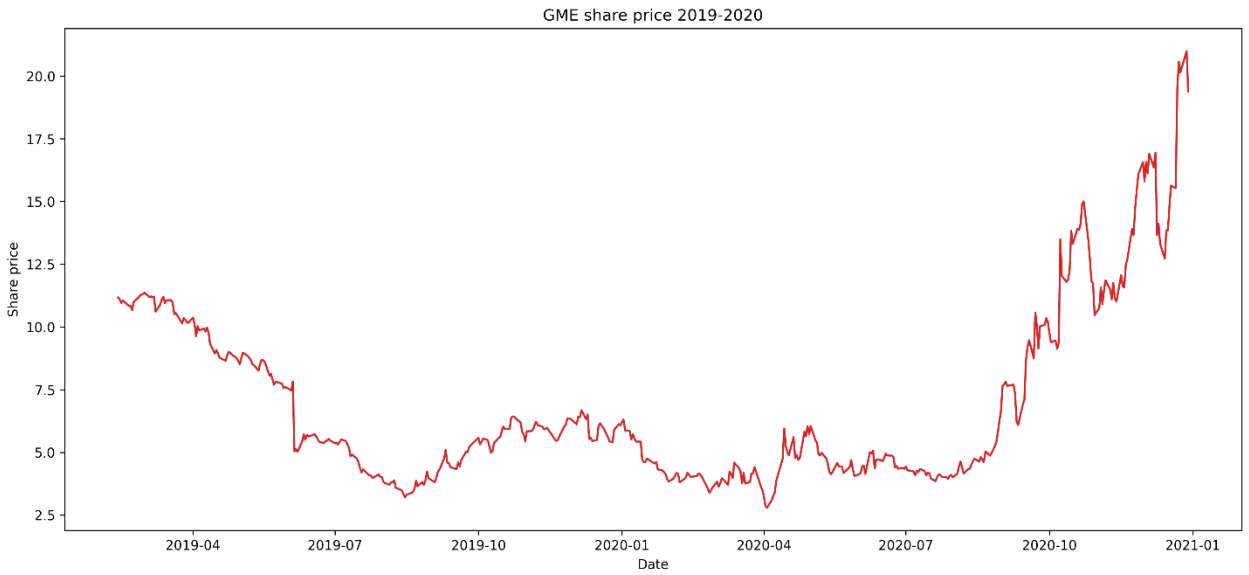


Figure 3: GameStop share price 2019-2020

On the subreddit of WallStreetBets, Gill's consistent monthly updates since 2019 on his GameStop returns became immensely popular and as a result, the stock gained traction of other retail investors and member of the subreddit. In the early stages of the incredible rise in stock price of GameStop, the subreddit's excitement was evident by the number of up votes on the GameStop related posts (Reddit, 2019). Gill personally explained in videos why he believed the rising stock value was justified. Later, multiple hedge funds had begun taking short positions on GameStop. They believed the stock was wildly overvalued, as stock prices soared to $20 during the Christmas period of 2020. GameStop had failed to report any profits since mid-2018, and the COVD-19 pandemic was still causing havoc.

In contrast to a regular long-holding investment fund, hedge funds can use short positions to increase their return on investment. When shorting a stock, the investor borrows and sells a stock on in the hopes of buying it back at a lower price (Vasileiou et al., 2021). If the share price depreciates, the investor's return is essentially the share price difference between the borrowed and the buy-back price. However, when share prices continues to rise, the investor must buy the stock back at a greater price than it originally borrowed it for, risking enormous losses. In the case of GameStop, several hedge funds took up shorting positions, such as Melvin Capital and Citron Research (Chung, 2021). This resulted in a challenge between the pessimistic hedge funds and the optimistic community of retail investors and member of WallStreetBets. GameStop's stock price kept on rising, influenced by retail investors and the phenomenon of short squeezing. A short squeeze happens when investors that shorted a stock sell their shorts to prevent further losses when share prices rise, resulting in an even higher share price (Vasileiou et al., 2021). It's a

self-reinforcing phenomenon. In the case of GameStop, the short squeeze outcome was a closing share price of $347 on January 27, 2020. The short squeeze resulted in some hedge funds defaulting due to their short positions in GameStop.

A small community taking on a fight with financial institutions and hedge funds was an unprecedented dynamic. Initially, the Wall Street Journal mentioned WallStreetBets on January 11, 2021, when it first became obvious that the subreddit's community had some sort of influence on the market performances of certain stocks (McCabe et al., 2021). Later, it appeared that the stock market behavior of GameStop influenced by a collective of retail investors was not a unique occurrence. Instead, a pattern showed of multiple stocks over performing due to its popularity on the subreddit of WallStreetBets.

2.1.6 Meme stocks

A meme stock, broadly defined, is a stock that gains popularity among retail investors through social media (The Economist, 2021b). Additionally, the popularity induced ascent characterizes a meme stock and is deemed by various stock market actors as irrational. They are also prone to being short squeezed as a reaction to overvaluation. Academic literature on the topic of meme stocks is lacking due to its recency. Instead, we used the most credible media sources available. GameStop is the original and most popular meme stock. The Wall Street Journal first mentioned the new phenomenon on the 11th of January 2021. They reported that retail investors rallied behind GameStop as an act of rebellion, intended to humble professional investors (McCabe et al., 2021). Other stocks such as AMC Entertainment also experienced similar large-scale traction on social media, especially on Reddit's WallStreetBets, sending its share prices to excessive valuations. Patterns of steep ascents in stock prices can be observed in all the meme stocks, driven by encouragement online, retail investors' fear of missing out and potential short squeezes. The popular posts of the ridiculous meme stock returns such as Keith Gill's, posted to the subreddit of WallStreetBets most likely encouraged other retail investors to take similar risks.

The most mentioned stocks on Reddit in January 2021 are AMC Entertainment Holdings Inc (AMC) and GameStop (GME). Among others frequently mentioned stocks are Tesla (TSLA), BlackBerry (BB) and Palantir (PLTR) (Ahmed, 2022). GameStop's activities have already been discussed extensively. The other four meme stock companies are briefly introduced in Table 2 by industry and activities. In this paper, we will try to analyze the popularity and sentiment of these five meme stocks on WallStreetBets and relate that to their stock market performance.

Company	Industry	Activities
AMC Entertainment	Entertainment	AMC Entertainment is a movie theatre company that owns and operates movie theatres in the US and Europe. (Corporate Profile | AMC Theatres, 2022)
Tesla	Automobiles	Tesla is a well-known electric vehicles manufacturer that also builds clean energy generation and storage products. (About Tesla | Tesla, 2022)
Palantir Technologies	Software	Palantir is a fast-growing US based software and data company that offers data solutions. Palantir attracts multiple government and military contracts in the US and Europe. (About Palantir, 2022)
BlackBerry	Software	Once one of the largest mobile phone manufacturers, BlackBerry now specializes in secure end-to-end communications solutions for the public and private sector. (BlackBerry Industry Solutions, 2022)

Table 2: Meme stock introduction

2.2 Evaluation stock market performances

A stock can only be a meme stock if it has gained popularity online, which consequently sends its share price to unforeseen valuations (The Economist, 2021b). As a result, the meme stocks' share prices are deemed to be overvalued by most market agents. In this section, we will explore ways to evaluate a stock's performance.

2.2.1 Price-earnings ratio (P/E)

The most widely used indicator of stock valuation and expected performance is the price to earnings (P/E) ratio (Ghaeli, 2017). The traditional price to earnings ratio is the ratio between the current share price and the earnings per share (EPS) of the previous year. Multiple variants of the P/E ratio can also be calculated with different timeframes or predicted future earnings.

$$\mathrm{P/E} = \frac{\text{Share Price}}{\text{Earnings per Share}}$$

Whilst the P/E ratio is the most common used evaluation method, interpretations can differ due to multiple factors (Anderson & Brooks, 2006). Four factors that have an influence on the P/E method of evaluation need to be discussed. The first being the year in which P/E is calculated. Average target P/E ratio varies year by year due to overall investors' confidence in the economy. The sector of the measured company can also be a big influence on the valuation of the P/E ratio. A fast-growing sector warrants a higher P/E ratio due to its expected earnings in the long-term. In contrast, firms in established, lower growth sectors

require lower P/E ratios to still be positively valued. The size of the company has a positive relation to the P/E ratio, as larger-capitalization companies tend to have higher share prices in comparison with its earnings. Lastly there is the idiosyncratic effect, meaning that company specific effects, such as positive news or the agreement of large contracts, can influence the evaluation of the P/E ratio. In conclusion, the P/E ratio is a popular and easy-to-use measure of evaluating stock performances yet requires scrutiny when comparing different companies in different sectors and sizes.

2.2.2 Enterprise value/EBITDA ratio (EV/EBITDA)

A second popular measure for comparing the valuation of a firm is the EV/EBITDA ratio. EV, the enterprise value, is the total equity value of the firm and is calculated by the addition of the firm's market capitalization (equity value) and its total debt. The EBITDA is the earnings before interest, taxes, depreciation, and amortization. This figure is often used as the firm's cash flow (Gupta, 2018).

$$\mathrm{EV/EBITDA} = \frac{\mathrm{equity} + \mathrm{debt}}{\mathrm{EBITDA}}$$

The EV/EBITDA ratio is a simple and frequently used metric for comparing the relative valuation of different companies. The ratio is reported as a multiple, such as 8x, meaning that the equity value of a company is 8 times the yearly cash flow or EBITDA.

2.2.3 Discounted cash flow model (DCF)

A different method of firm valuation is the discounted cash flow (DCF) model. The DCF is a projection based on the expected future cash flow derived from historic data (Kruschwitz & Loeffler, 2005). It values the equity of a firm by discounting free cash flow from the firm's operations to present values subtracted by the value of the firm's debt. As such, the DCF model attempts to valuate a company at present based on the projected cash flow in the future. The valuation of the company divided by the outstanding shares results in the estimated valuation of its share price. Next, we will discuss the model's formula and its components.

$$\mathrm{DCF} = \frac{\mathrm{CF}_1}{1 + \mathrm{r}^1} + \frac{\mathrm{CF}_2}{1 + \mathrm{r}^2} + \cdots + \frac{\mathrm{CF}_\mathrm{n}}{1 + \mathrm{r}^\mathrm{n}}$$

where: CF = cash flow, CF_1 being the cash flow of the first year
r = discount rate

Cash flow, or Free Cash Flow (FCF) can either be free cash flow to the firm (FCFF) or free cash flow to equity (FCFE). FCFF is the free cash flow available from operations after paying all expenses, without considering interest and debt repayment. FCFE is the free cash flow available to the equity shareholders of the firm after accounting for all debts, expenses, and reinvestment. In this model, FCFE is preferred as it is a more encompassing measure to calculate a firm's realistic valuation (Kruschwitz & Loeffler, 2005).

The discount rate is used to correct depreciation over time to calculate a present value of a future amount. In its most straightforward use, the discount rate is the annual interest rate of an investment. When used to estimate the net present value of a company, the discount rate is substituted by the Weighted Average Cost of Capital (WACC), the blended cost of the firm's capital from every source (Kruschwitz & Loeffler, 2005).

Weighted Average Cost of Capital (WACC) is the addition of the weighted cost of debt and equity of the firm. The data for the components can be easily found on the firm's public balance sheet. However, the cost of equity and debt require another calculation (Kruschwitz & Loeffler, 2005).

$$\mathrm{WACC} = \left(\frac{\mathrm{E}}{\mathrm{V}} \times \mathrm{R_e}\right) + \left(\left(\frac{\mathrm{D}}{\mathrm{V}} \times \mathrm{R_d}\right) \times (1 - \mathrm{T})\right)$$

where: E = market value of firm's equity, market capitalization
D = market value of firm's debt
V = total valuation of debt and equity
R_e = cost of equity
R_d = cost of debt
T = capital tax rate

Cost of equity is the opportunity cost of capital. Investors require a certain return rate to compensate the risk that involves in the investment (Kruschwitz & Loeffler, 2005).

$$\mathrm{R_e} = \mathrm{R_f} + \beta \times (\mathrm{R_m} - \mathrm{R_f})$$

where: R_e = cost of equity
R_f = the risk-free rate, usually the 10-year U.S. treasury bond
β = the relative risk of the stock
R_m = the annual return of the market

The risk-free rate is usually the return rate of a 10-year U.S. treasury bond. The treasury bonds are used for safe investments. The beta of a stock is the relative volatility or riskiness when compared to all other stocks in the market. This data can be readily found on Bloomberg. Lastly, the annual return of the market is the historic average return rate that stock market delivers. Usually, the return rate of the S&P500 is employed.

Cost of debt can simply be calculated by dividing the interest expenses from the addition of short- and long-term debt of the company. The interest expenses are usually tax-deductible. To correct this, the cost of debt is multiplied by one divided by the capital tax rate.

The discounted cash flow model does have its limitations due to the assumptions of certain costs, values for the annual return of the market and the relative risk of the stock. However, it remains a popular and useful model for firm valuation. When the DCF valuation is divided by the outstanding shares, conclusions can be made on the valuation of the firm's current share price (Kruschwitz & Loeffler, 2005).

2.3 Data Mining and Machine Learning

Following the macroeconomic context and stock valuation methods, we will now introduce the concepts of data mining and machine learning. A standard for data mining processes is adopted for this research's methodology, using the appropriate machine learning techniques.

2.3.1 Data mining process

Data mining is a process of extracting information from data, organized by multiple stages. (Provost & Fawcett, 2013). The process starts by having a clear understanding of the business problem whereby a data mining solution could be an answer. An effective data mining process requires the formulation of clear instructions and consultation with important stakeholders (Brynjolfsson, Hitt & Kim, 2011). The next step to consider is different machine learning techniques to efficiently solve the business problem. Machine learning (ML) is a method of data analysis that automates analytical model building (Sedkaoui, 2018). Knowing which types of machine learning techniques are appropriate to solve a problem is an invaluable skill for data scientists (Provost & Fawcett, 2013). The most common machine learning techniques that solve problems with algorithms are classification, regression analysis and similarity matching. These techniques can be classified by the amount of human supervision the ML models require.

2.3.2 Supervised, semi-supervised and unsupervised machine learning

There are three methods of building a machine learning model (Van Engelen & Hoos, 2019). Traditionally, supervised- and unsupervised learning are the two main methods of machine learning. A machine learning model is using supervised machine learning when a preconceived target is defined. Classifying cats and dogs by image recognition serves as an example of supervised learning. The supervised ML model already has predefined categories. Regression analysis and classification are commonly used methods of supervised learning. In contrast, machine learning with unsupervised learning do not have any set objectives (Längkvist, Karlsson & Loutfi, 2014). The unsupervised ML model analyses the data and categorizes certain data points, possibly resulting in new insights. However, the resulting classification is not necessarily significant. Recently, the hybrid method of semi-supervised learning was introduced to cope with machine learning that have large sized and disorganized datasets (Van Engelen & Hoos, 2019). Semi-supervision is only applied to complex and large unsupervised machine learning models. These models have the potential to perform more accurately when using supervised learning on limited segments of the dataset as a sort of guidance. In any data mining process, it is essential to establish which kind of machine learning method is required.

2.3.3 CRISP-DM

The data mining process serves as a structured methodology for solving machine learning problems (Shearer, 2000). Using a universally accepted methodology results in higher consistency and reproducibility. One of the methodologies for machine learning problems is the cross industry standard process for data mining, CRISP-DM (Shearer, 2000). According to Schröer et al. (2021) CRISP-DM is the most professionally used data mining process. The visualization of the industry standard is displayed in Figure 4. Shearer visually uses interactive arrows and a loop, emphasizing the continuous iterative nature of data mining. Six steps need to be followed to complete the process.

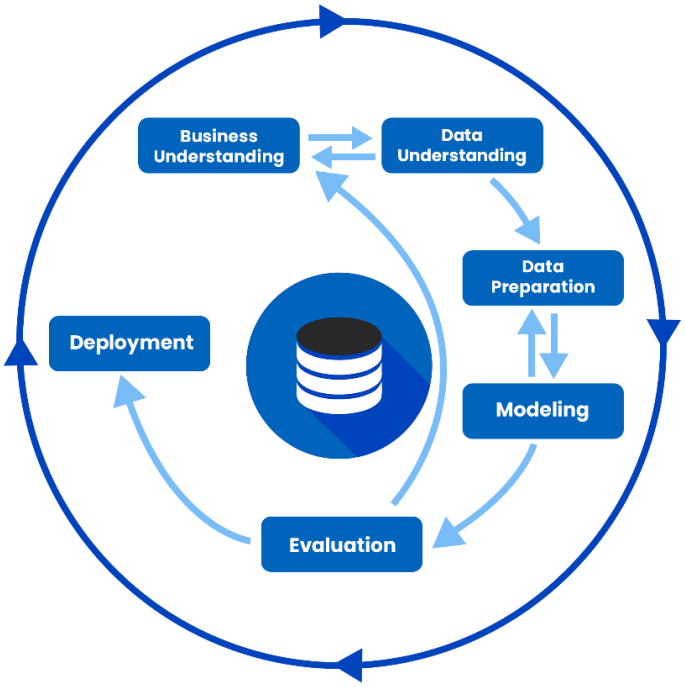


Figure 4: CRISP-DM

Business understanding is the crucial first step of CRISP-DM. Clearly formulating and visualizing the business problem is essential for an effective data mining process (Brynjolfsson, Hitt & Kim, 2011). Stakeholders must define the requirements of the data mining solution. Due to the iterative nature of the process, business problems and use-case can be reformulated as an outcome of a deeper data understanding or evaluation.

Data understanding is the second phase of the CRISP-DM model. The data used in the data mining process critically influences the machine learning model's usability. In this step, choices need to be made about the use and acquisition of datasets. The dataset should be in line with the machine learning model that tries to solve the formulated business problem. Additionally, the following aspects of the dataset should be considered: cost, accuracy, availability, and the strengths and weaknesses of each dataset (Provost & Fawcett, 2013). In the end, the performance of the ML model depends on the quality of the dataset that it has been performing on.

Data preparation is the most time-consuming aspect of machine learning. In this phase, the datasets decided upon in the data understanding step of CRISP-DM are structured and adjusted to enhance the performance of the ML model. Corrections are made and unusable data points are taken out. Data preparation is also an iterative process depending on the modelling phase and the possible new datasets that become available during the process.

Modelling is the implementation of a machine learning model to solve the business problem. During the CRISP-DM process, multiple versions of the ML models are built to repeatedly try to increase its performances. The stage closely links back to the data preparation phase. The difficulties of model building predominantly depend on the quality and structure of the data (Shearer, 2000).

Evaluation is the decisive phase in the CRISP-DM process where the performance and usability of the ML model is evaluated in regard to the business problem (Schröer, Kruse, & Gómez, 2021). Stakeholder consensus must be reached to further progress to the implementation phase. It is essential that the ML model poses a valid solution to the formulated business problem. Questions should be asked about the choices made in datasets and machine learning techniques used. Additionally, different methods for evaluation should report about the performance and usability of the ML model. These evaluating methods will also be discussed separately. If the evaluation phase is unfavorable, the process returns to the business understanding phase with valuable lessons learned in the previous iteration. If the performances are favorable, then the model is implemented in the last phase.

Implementation is the final phase of CRISP-DM. The carefully evaluated model is ready to be launched and used. The desired output depends on the type of model and business problem. An intelligent system deployed to a business process is a typical example of implementing a machine learning model. Another type of output is simply gaining information (Provost & Fawcett, 2013).

The six-step process of CRISP-DM offers a repeatable and consistent methodology for data mining solutions, ensuing its popularity in the industry (Schröer et al., 2021). Therefore, it is also our preferred method for structuring this research.

2.4 Python

Before we introduce the techniques used in this research, we will first define the environment in which they are built. Python is an open-source programming language and is used for every aspect of coding in this research. It is especially fast and powerful, while also being very linkable with different applications (Van Rossum, 1991). Python hosts an enormous number of third-party modules or libraries, which are chunks of reusable codes. In this research, numerous Python libraries are used for processing data, machine learning and deep learning.

Many options exist for Python coding interfaces. Jupyter notebook is an easy-to-use interface for Python coding (Jupyter Notebook, 2014). It is a simple web application that can effortlessly share documentation and produce visualizations. This makes it our interface of choice. Data visualization is important for data science as it makes data understanding easier (Unwin, 2020).

2.5 Web scraping

Web scraping is the practice of extracting data from websites (Mitchell, 2018). It is accomplished by writing software to extract the textual elements needed for data collection. Those textual elements can be found in the structured code of websites. Web scraping can be performed manually or automatically by programming (Emerson, 2019). It is widely used for research and business purposes, such as online sentiment analysis. In our case, we will gather as much data as possible from Reddit's WallStreetBets. Free software is widely available to scrape websites online. We will use pushshift.io, an application programming interface or API for Python, to read the posts of Reddit.

2.6 Natural Language Processing (NLP)

Natural language processing (NLP) is a subdomain of computer science and artificial intelligence. In general, NLP is concerned with computer understanding of human language. Many fields exist within NLP (Nadkarni et al., 2011). In our research, the NLP models used are sentiment analysis and rule-based matching. First, we will introduce the underlying functioning of sentiment analyzer models.

2.6.1 Approaching text analysis through supervised learning

Historically, the default approach to sentiment analysis has been through supervised learning algorithms. Support Vector Machine was used as a method for early polarity classifications together with a Bag-of-Words model (Paltoglou & Thelwall, 2010).

Support Vector Machine (SVM) is a supervised learning algorithm method. The most straightforward use of SVM's is in a two-dimensional space. The SVM aims to separate two categories to the best of its ability by using historic data (Ng, 2018).

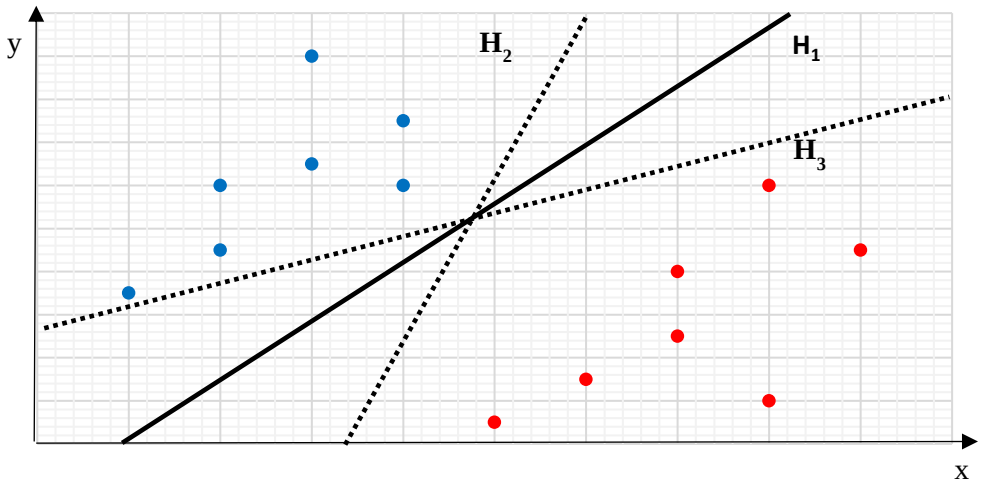


Figure 5: Support Vector Machines

In Figure 5, the three lines each separate two classes of instances, represented by the colored dots. The SVM finds the optimal classification by fitting a line that averages the greatest distance from the closest dots or instances. In this case, the H_1 line separates the two classes the best. H_2 and H_3 both have instances close to its separation line. The instances used in Figure 5 were already categorized to train the SVM. When a trained SVM becomes operational, a newly added instance can be classified to a class. More complex versions exist of SVM's that include the addition of multiple classes and/or multiple dimensions. These are out of the scope for this paper.

A **Bag-of-Words** (BOW) is used to gather an unordered collection of words in a document. The relative frequencies of the words present in a document provide a baseline classification for sentiment analysis (Pang et al., 2002). An example of the BOW approach could be the following:

(1) Retail investors are using Robinhood to jump into option trading.
(2) Robinhood started trading on the Nasdaq stock exchange.

These two sentences are gathered in a Bag-of-Words, resulting in an unorganized output of the words and their relative frequencies.

Bag-of-Words = {'Retail': 1, 'investors': 1, 'are': 1, 'using': 1, 'Robinhood': 2, 'to': 1, 'jump': 1, 'into': 1, 'option': 1, 'trading': 2, 'started': 1, 'on': 1, 'the': 1, 'Nasdaq': 1, 'stock': 1, 'exchange': 1}

Term frequency can be a valuable measure for analyzing a text. There are some issues however with the naïve collection of terms. A word that is used twice or more is counted independently if one has an uppercase letter. Similarly, spelling mistakes pose problems. A possible solution to these problems is to normalize the data. Normalization means converting text into the same standard lowercase form. 'Canada', 'Canada', 'CanadA' and 'CANADA' are normalized to 'canada' allowing for more accurate and consistent processing. Other useful steps to process textual data include removing stop words and removing non-alphanumeric characters. Stop words are terms that do not have any meaning, such as 'the' or 'a'. While a non-alphanumeric character is any character that is not a numeric or alphabetic character, such as an underscore or percent sign.

Term frequency–inverse document frequency (TFIDF) is a measure for attributing importance to a term in a certain document (Ko, 2012). In the English language, there is a high frequency of the word 'the'. As a result, the relative importance attributed to 'the' should be small. TFIDF uses the inverse term frequency to attribute relative importance or weight to the term. A scarcely used term in a certain text or document will result in a large importance. The formula for the TFIDF is as follows:

$$w_i = tf_i \,.\, idf_i = tf_i \,.\, \log \frac{N}{df_i}$$

where: w_i = weight of a term i.

tf_i= the number of times term i occurs in the document D

idf_i = the inverse document frequency of term i.

N = the number of documents

df_i = the number of documents containing term i.

2.6.2 Sentiment analysis and text classification

Sentiment analysis is a field of natural language processing that deals with the computational treatment of opinion, sentiment, and subjectivity in a text (Pang & Lee, 2008). The categorization of sentiment is called the text classification. A machine learning model can be trained to identify semantic polarity in a text. Traditionally, semantic polarity has been used for predicting elections and movie sales (Jung-Tae Jo & SangHyunChoi, 2015). The strengths of semantic analysis models are especially apparent in the application of information gathering on social media like Reddit and Twitter, where users are urged to express their opinion. Wright (2009) already identified the economic value of online opinion in 2009, stating that it had turned into a kind of virtual currency that can make or break a product. There are various options available for sentiment analysis, two are used in this research and will be discussed later.

In our research, sentiment analysis evaluates subjectivity exclusively in textual format. A major difficulty with polarity classification is ranking sentiment on a scale. Many sentiment analysis models use SVM to rank and classify. Ranking sentiment is especially difficult when the polarity becomes less obvious (Chunxi Liu et al., 2011). This can often occur during elections sentiment analysis, where being supportive or being opposed to policies or politicians are not clearly defined polarities. Another difficulty with sentiment analysis is context. Depending on many contextual aspects, an ordinary statement could be expressed sarcastically or not, which is tough to detect.

In this case, sentiment analysis is performed on the subreddit WallStreetBets. The subreddit or forum has a unique vocabulary where emojis are frequently used to express sentiment. ‘GME to the moon’ followed by a moon or rocket emoji is an example of frequently used terminology. Objectively, this sentence does not make much sense. However, in the context of WallStreetBets, this expression is often used to express very positive sentiment. The vocabulary or lingo of the subreddit is an ever-changing entity riddled with jokes and sarcasm. This is especially tough for sentiment analysis (Mohammad, 2016). The subreddit also has an extensive meme culture which has been influencing other parts of Reddit. The memes posted on WallStreetBets often require context and can seem so absurd at times, to then disappear in a matter of days.

2.6.3 TextBlob sentiment analysis model

TextBlob is one of two sentiment analysis models used in this research. It returns the average sentiment from a text by scanning positive and negative words (Loria, 2020). The positivity and negativity of the words are recognized by another software library called Natural Language ToolKit (NLTK). Their sentiment analyzer has been trained on book and movie reviews (NLTK, 2001). NLTK ranks the sentiment from negative to positive. For this sentiment analysis model, additional training is not necessary as it has already been trained on English book and movie reviews with a resulting Bag of Words.

2.6.4 Scikit-learn SVM sentiment analysis model

Scikit-learn (sklearn) SVM sentiment analysis is the second sentiment analysis model used. Sklearn is an extensive library made for computer science on Python (Scikit-learn, 2007). Compared to TextBlob, there are many more options for building the sentiment analysis model. Any kind of compatible tokenizer or classifier can be implemented. In this case, we will use the traditional support vector machine algorithm as a classifier (Paltoglou & Thelwall, 2010). The sklearn model requires some training data, on which the model can fit itself to. Because of the specific sentiment analysis on the WallStreetBets subreddit, the training data is a manual and laborious task of categorizing the posts into positive or negative sentiment.

Sklearn uses a pipeline feature for transforming and estimating data. There are three components in our pipeline. First, our textual data is lowered and prepared in the cleaner component of the pipeline. Next, the SpaCy tokenizer is used to split up the posts into individual words or tokens and remove stop words, punctuation and pronouns in the process. The last component of the model is the estimator or predictor, which is a support vector machine that classifies the words into sentiment categories. The model is trained by fitting the training data. To evaluate this model, we predict the training data using the pipeline and compare the predictions against the actual sentiment that was manually annotated.

2.6.5 Rule-based matcher

Rule-based matching is a useful tool for classifying texts by a custom linguistic rule set (SpaCy.io 2021). Any combination of linguistic rules can be set in the matcher. For example, one defined rule could be the combination of (punctuation = true) and (lowercase = 'gme'). The matcher will recognize any combination with punctuality followed by the text 'gme'. As a result, the matcher will find both '.GME' and '?gme' in each text and classify them in the defined category. Any possible grammatical rule can be defined by the matcher. Its flexibility for finding specific forms of text in a document makes it a powerful tool for NLP. In this research, the rule-based matcher from SpaCy.io is used to find all possible mentions of meme stocks in the data.

2.7 Linear Regression

Linear Regression is a supervised learning algorithm. It is used to predict an output within a continuous range (Chen, 2020). Both simple (one independent variable) and multiple (many independent variables) linear regression exist. In this research, only multiple linear regression is used.

Multiple linear regression (MLR) (Chen, 2020)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

where: y = dependent variable

β_0= the intercept

$\beta_1, \ldots, \beta_n$ = the coefficients of the independent variables

$x_1, \ldots, x_n$= the independent variables

ε = random error

The output of linear regression models is twofold (Chen, 2020). One desired outcome of linear regression is the accurate prediction of a dependent variable by an input of independent variables. The other aim of linear regression is to analyze the predictive nature of a specific independent variable. In this research, both outcomes of linear regression are relevant. Predicting the share price of a meme stock using dependent variables and analyzing the information gained by each of those variables.

2.8 Neural Networks (NN)

In our research, we will use neural networks to predict meme stock prices. Neural networks are computing systems that try to recognize intricate structure in large datasets (LeCun et al., 2015). They are used as algorithms in the subdomain of machine learning called deep learning. To better comprehend the inner workings of neural networks, we will briefly introduce the necessary foundations of neural networks and recurrent neural networks.

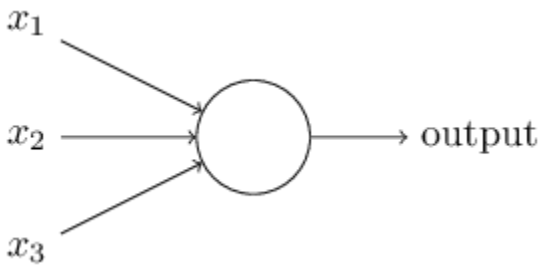


Figure 6: Sigmoid neuron (Nielsen, 2015)

Neural networks are a network of layers consisting of neurons, that take numerical values as input and produce a single output (Nielsen, 2015). In Figure 6, a single artificial neuron is depicted, called a sigmoid neuron. It has three inputs (x_1, x_2, x_3). These inputs each have a numerical value x between 0 and 1, a weight w (w_1, w_2, w_3) and a bias b. When the inputs are received by the neuron, it computes the sum of all these inputs. These are the numerical values multiplied by their respective weights with the addition of the bias. The output of the sigmoid neuron takes the sum of the inputs and calculates a sigmoid function, $\sigma(\sum w_i . x_i + b)$.

$$output = \frac{1}{1 + e^{(-\sum w_i . x_i - b)}}$$

The output of the sigmoid function only has values between 0 and 1. When the sum of the inputs has a high positive weight and bias, the output will approximate 1. This is due to the minus sign in the exponent, which causes a rapid descent to zero with high positive values. In contrast, when the inputs are very negative, the output will converge to 0 because the exponent function becomes positive. The denominator becomes very large, and the resulting fraction approximates zero. Crucial to the sigmoid neuron's output is the weights of the inputs (Nielsen, 2015). The output could change on slight adjustments on the weights of the inputs.

Neural networks consist of many layers of neurons called hidden layers. In Figure 7, there are two hidden layers consisting of one or many neurons. The output of the first layer becomes the input to the second layer. Every one of the neurons in the neural network individually calculates an output and sends it through to the next layer. Each layer adds complexity to the network.

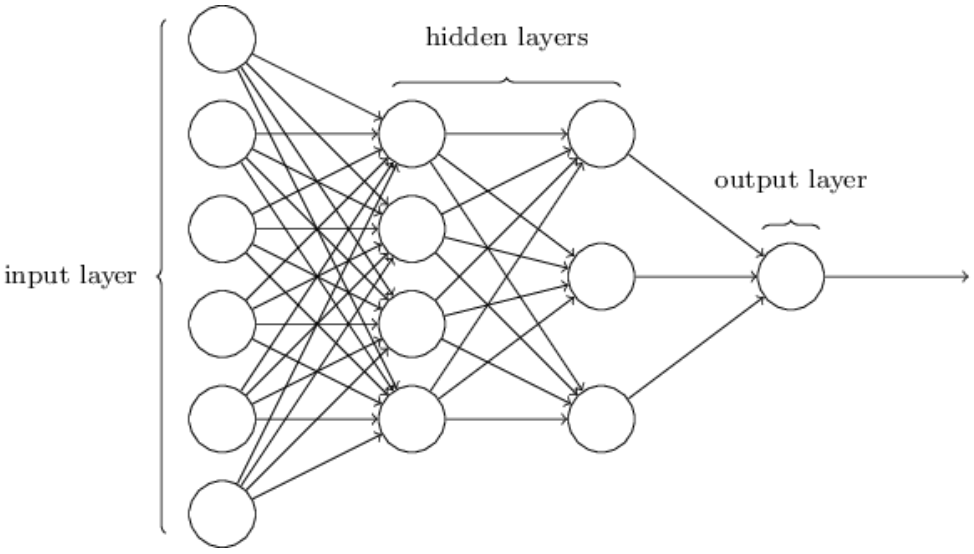


Figure 7: Simple Neural Network. (Nielsen, 2015)

Neural networks train iteratively by changing weights using backpropagation (Goodfellow et al., 2016). Backpropagation is an algorithm that is used for calculating the gradient of an error function of the weights used in a neural network. The algorithm works its way backwards in the many steps of a neural network in search of partial derivatives of the error with respect to the weights of each step. This enables the network to adjust these errors (Nabi, 2021). In essence, backpropagation searches for errors in the weights and gives back a value. Later, gradient descent is introduced. This other algorithm can iteratively minimize values for a given function. Gradient descent is used in backpropagation for adjusting the weights up or down of a neural network, which is how they train.

2.8.1 Recurrent Neural Networks (RNN)

Recurrent neural networks (RNN) are a type of neural networks used for processing sequential data ($x_t = \mathrm{x}_1, \cdots, \mathrm{x}_n$) where previous inputs are used as inputs for predicting a current output (Amidi & Standford.Edu, 2019). Sequential inputs, such as typing words into a search bar one after another, can be predicted by using RNN. These networks are recurrent since they perform the same task of predicting an outcome for any step in the sequence, and because they are dependent on previous input. In essence, RNN 'remember' information from previous steps and, as a result, are good at predicting.

Figure 8 shows an RNN that is being unfolded into its layers (Goodfellow et al., 2016). The network has the same number of layers as the length of the sequential data. As an example, a five-word sentence has five layers in an RNN. Next, we will go over each aspect of the displayed network.

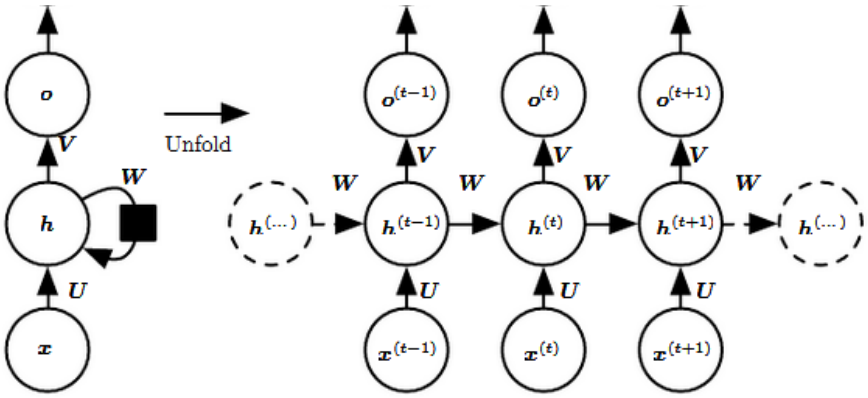


Figure 8: RNN (Goodfellow et al., 2016)

Input: $x^{(t)}$ is the input of the network, 't' stands for the time step and is situated in Figure 8 as the middle input in the unfolded network. In our example, an input could be a word from a sentence.

Hidden state: $h^{(t)}$ is the 'memory' part of the neural network. Neural Networks that try to predict the future from the recent hidden states. A hidden state of a current time step is a function of the previous step $h^{(t-1)}$ and from the current input $x^{(t)}$ corrected by a weight. This function can be any sort of transformation.

$$h^{(t)} = f(h^{(t-1)}, x^t)$$

Weights: U, W & V are weights matrices for the input, hidden state and output. They parametrically characterize connections of inputs and outputs by weight. In other words, they can influence the input-to-hidden, hidden-to-hidden and hidden-to-output connections by a multiple.

Output: $o^{(t)}$ is the output of the network at time step t.

While recurrent neural networks have amazing possibilities and strengths, there are some problems with the gradients. Gradients are partial derivatives of an input (Nabi, 2021). They measure how much the outcome of a function changes given an input. The gradients of inputs are used to train a neural network. The outcome of the neural network can be altered by changing the inputs and their weights, which can be tracked by the gradients. The problem that exists with RNN's is the vanishing gradient problem (Hochreiter, 1998). When training a neural network, the gradients can become so small, that the updates given to the model by backpropagation are ineffective. These updates depend on the gradients, and as a consequence, the model stops learning altogether. Since then, new types of neural networks have been created to overcome this issue, such as the Long Short-Term Memory (LSTM) or Gated Recurrent Unit.

2.8.2 Gated Recurrent Unit (GRU)

The Gated Recurrent Unit is an upgrade to the recurrent neural network introduced by Cho et al.(2014) aiming to solve the vanishing gradient problem. The solution was to add additional complexity to the RNN so that it can keep relevant historic information from far back in the network. In each hidden state of a GRU, there is an update and reset gate that decides which information should be kept and sent to the output. The inner mathematical workings of the GRU are out of the scope for this research and will not be discussed. Instead, we will focus on the implementation of the GRU on our research case in a Python environment.

GRU is used as our recurrent neural network of choice due to the fast performance when compared to others such as the long short-term memory LSTM (Mateus et al., 2021). Tensorflow, a library for Python, provides the ability to build GRU's from the ground up with an API from Keras. In our case, the GRU will predict the stock market price of the meme stocks one step ahead, meaning the next trading day. This means that the dependent variable is the closing share price of the meme stocks. The complexity of the GRU is also adjustable. We will use multi-variate forecasting, meaning that the multiple independent variables are used, and additional layers will be added.

2.9 Machine Learning evaluation

Various machine learning models require different evaluation techniques. We will now examine the evaluation measures used in this research.

2.9.1 Confusion matrix and indicators

The confusion matrix is a practical tool for evaluating machine learning models (Provost & Fawcett, 2013). It offers insights into the performance and the shortcomings of the model. The confusion matrix demonstrates the model's predictive behavior when classifying data into positive and negative values. The predictions are compared to the actual values as evaluation.

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive** **Type 1 error**
	Negative	**False Negative** **Type 2 error**	**True Negative**

Table 3: Confusion Matrix (Provost & Fawcett, 2013)

Where**:**

TP: true positive, correctly predicted as positive.

TN: true negative, correctly predicted as negative.

FP: false positive, incorrectly predicted as positive. Should be negative.

FN: false negative, incorrectly predicted as negative. Should be positive.

The **precision** of a model is defined as the ratio between the true positives and the sum of true positives and false positives. A high value for precision indicates that the model accurately predicts nearly all the true data points. Consequently, there should also be a relatively low number of false positives detected. The model does not produce many type-1 errors (Martínez Torres et al., 2019).

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

Another measure for the evaluation of the model is the **recall**. This is the ratio between the true positives and the sum of true positives and false negatives. A high value for recall indicates that the model is great at classifying true positives without having many incorrectly predicted negative data points (type 2 error). A low value of recall means that the model predicts a considerable number of true data points as negative. This might be an indication that the model has too many strict guidelines for classifying a data point as true.

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

The weighted average of the previous measures is called the **F-score**. This is a great way to evaluate the overall usability of the model. However, breaking the F-score down to both recall and precision is essential to understand what kind of errors the model has difficulties with.

$$Fscore = \frac{Recall + Precision}{2}$$

2.9.2 Overfitting

Overfitting is the tendency for a machine learning model to perform remarkably well on the dataset it has trained on (Karystinos & Pados, 2000). On the other hand, an overfitted model lacks performance when deployed on a new dataset. This especially becomes an issue when the machine learning solution is deployed to a business process that continuously feeds new data to the model. Overfitting is inherent to building ML models, as it learns from the provided data (Provost & Fawcett, 2013). It is necessary to combat overfitting by implementing multiple precautions. The most effective measure is splitting the data used to train the model. In this way, the dataset is divided into a training dataset and a validation dataset. The machine learning model learns from the training dataset and is continuously evaluated by the validation dataset.

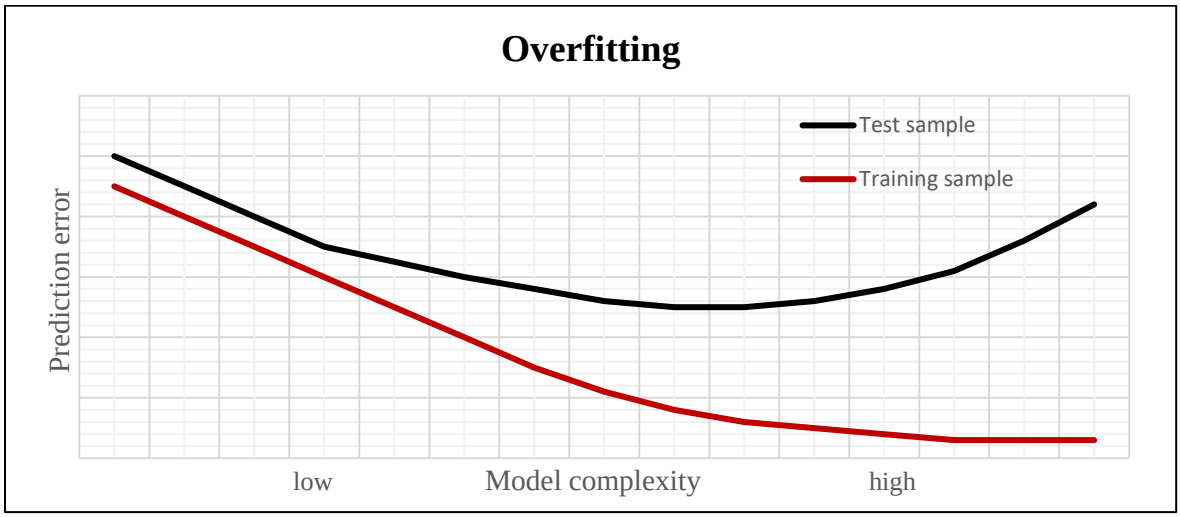


Figure 9: Overfitting

In Figure 9, the prediction error rate is plotted against the model complexity. The model complexity increases when the model trains iteratively more on a given dataset sample. The performance of both the test and training sample are identified. However, only the training sample is used to train the model. Initially, the model performs well on both the test and training sample when the model's complexity is still low. When the model complexity develops, the performance on the training sample still improves. On the other hand, the improvements of the prediction error on the test sample start to stagnate. At a certain optimal complexity, the test sample's prediction error is at a minimum. Eventually, the prediction error starts to increase again, which indicates that the model has begun to overfit to the training sample (Provost & Fawcett, 2013).

2.9.3 Linear Regression and GRU Analysis

There are many aspects to evaluating multilinear regression. There are indicators for the errors and for the effectiveness of the model and its components. We will discuss them one by one.

R^2, or R-squared, is the proportion of variance for the dependent variable that is explained by the model and its independent variable. The R^2 put simply is a percentage of how well the model's inputs explain the result of the regression analysis. R-squared is used to measure the strength of the model (Provost & Fawcett, 2013).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where: SS_{res} = the sum of squares of residuals

SS_{tot} = the total sum of squares

The residual sum of squares measures the squared overall difference between the predicted values, and the actual values while the total sum of squares is the squared difference between the dependent variable, (closing share price) and its mean. It strictly used for linear models (Spiess & Neumeyer, 2010).

MSE or mean squared error is a common metric for errors on regression analysis. The average is taken from the squared difference between the predicted value and its actual value for all its predicted values. Due to its squared nature, MSE heavily punishes outliers.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$$

where: n = number of predicted values

y_i = predicted value

$\widehat{y_i}$ = actual value

MAE or mean absolute error is similar to the MSE but does not square the difference between the actual values and the predicted ones. It is therefore a more forgiving metric for outliers in the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i|$$

where: n = number of predicted values

y_i = predicted value

$\widehat{y_i}$ = actual value

RMSE or root mean squared error is the square root of the MSE. It is a measure of accuracy for forecasting models. RMSE is used to compare regression models' effectiveness (Hyndman & Koehler, 2006). Outliers have a large effect on RMSE's performance due to squaring errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}$$

where: n = number of predicted values

y_i = predicted value

$\widehat{y_i}$ = actual value

An effective regression model should aim to have low values of MSE, MAE and RMSE. A low MSE means that the model accurately predicts values and consequently does not predict many outliers. R^2 should be as high as possible. An R-squared value of 1 would mean that the independent variables explain the models entirely (Provost & Fawcett, 2013).

The coefficients in multiple linear regression describe the relationship between the specific dependent and independent variable. The sign of the coefficient indicates the direction of the relationship, while the absolute value of the coefficient is the weight of a one unit change in the dependent variable.

3. Methodology

In this chapter, we will discuss the methodology used to answer our research question. Soni et al.(2022) systematically reviewed machine learning approaches to stock price predictions. According to their review, a valid method for stock price prediction is the combination of sentiment analysis of a stock in combination with numeric historic stock data. In our approach, we will set up a sentiment analysis on posts of WallStreetBets (WSB) and combine this gathered information to the numerical historical values of each meme stock using multiple linear regression and gated recurrent unit.

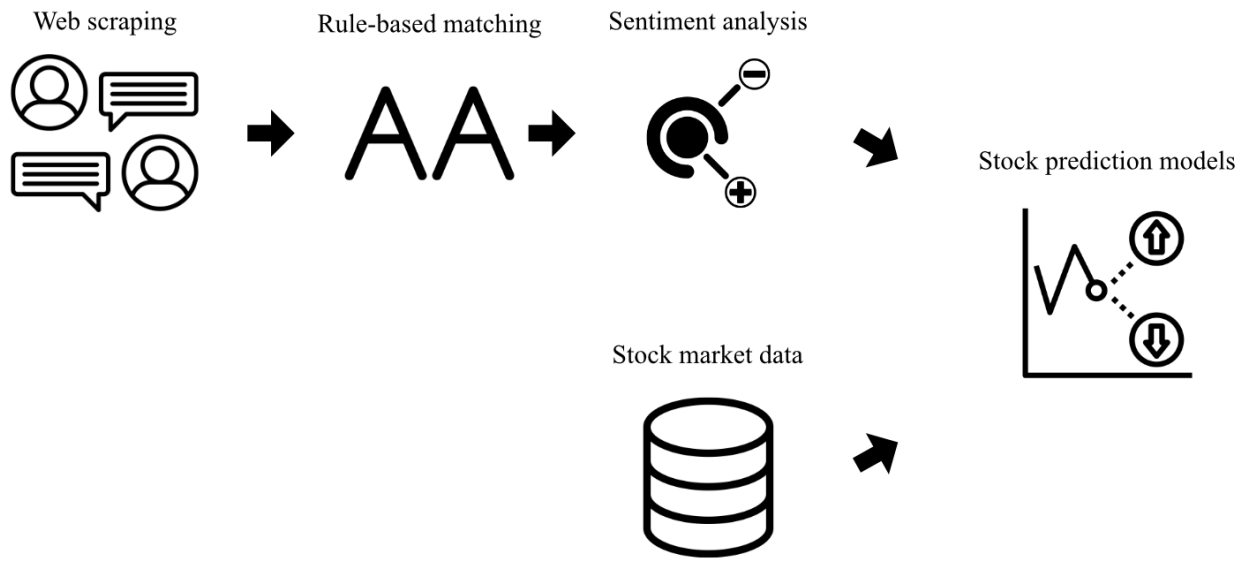


Figure 10: Overview of research process

A brief chronological overview of the research process is presented in Figure 10, from data gathering to stock prediction models. First, the WallStreetBets data is retrieved by scraping the subreddit, followed by rule-based matching that categorizes the WallStreetBets data by mention of meme stock. Next, the WallStreetBets data is evaluated by a trained sentiment analysis model. Its output is merged with stock market data. The combined dataset serves as input to the two stock prediction models, multiple linear regression and gated recurrent unit. The CRISP-DM process is followed throughout the course of the data mining case. Each of the steps will be discussed in detail.

3.1 Business understanding

The aim of this research is to predict and acquire insights into the effect that the joint market action of retail investors on WallStreetBets has on the most popular meme stock prices. Data from WallStreetBets is used in combination with stock market date for the purpose of linking the meme stock sentiment of the

Reddit community with the meme stock's performance. If any correlation exists between these two, the resulting machine learning models can hopefully predict at least some of the meme stock's market behavior. A possible use-case for the stock price prediction models could be the timely indication of an opportunistic entry point for investments in meme stocks, as well as identifying the even more interesting exit window. The prediction of share prices using sentiment analysis is just one of the possible outcomes of this research. Reddit's popularity as a social media platform is still relatively new and growing fast. The Reddit community of WallStreetBets has been underestimated in its market power, resulting in multiple defaults for market agents. Any new information on this subreddit's behavior is valuable, as research is lacking on this topic.

To add more context to the business understanding of this data mining case, it is valuable to look at the meme stock share price valuations at their peaks. In the result section of this research paper, we will discuss the extent of the meme stock price overvaluations. The three methods for evaluating the share price are the price to earnings (P/E) ratio, the EV/EBITDA ratio, and the discounted cash flow (DCF) model.

3.2 Data understanding

The data used for analyzing the sentiment and hype of the WallStreetBets subreddit is scraped off the Reddit online platform. There are multiple libraries on Python that allow data to be extracted from webpages using HTML elements. We will use the API of pushshift.io to retrieve all the relevant data. The web scraping procedure collects multiple elements such as the title, date, score and number of comments of all the posts between 2020 and 2022. A high score (net number of upvotes) and number of comments of the post indicates that the post could be more important or controversial than a post with a lower score or number of comments. Only the titles are used for the sentiment analysis in this research, while the other collected data is later used as independent variables for the prediction models.

Title	**Score**	**#Comments**	**Timestamp**
How do I buy AMC/GME stock?	2	21	4/02/2021 14:45
Remember HODL (not financial advice)	972	46	19/03/2021 1:02
What if; A catastrophic space expedition	5	51	11/07/2021 21:07
I don't care if I lose every cent, I'm holding	1	1	29/01/2021 1:57
BB about to 🚀🚀🚀 Dip being bought	5	0	29/01/2021 1:20
AMC To the moon?	42	22	5/02/2021 4:02

Table 4: Sample WallStreetBets data

The stock market data is obtained from Yahoo Finance with the Python library Yfinance. This library can be called on a Python environment to retrieve share price data for any stock during any period. In this research, five meme stocks are analyzed. GameStop, AMC Entertainment, BlackBerry, Palantir Technologies and Tesla. We retrieved stock market data for each meme stock between 2020 and 2022. Some extra stock data from 2019 was called to calculate the 30-day trailing moving averages. The stock data include the daily volume, price of opening, daily (adjusted) close, daily high and daily low. The Yfinance stock market dataset is clean and accurate, which is ideal as an input for the prediction models.

Date	Open	High	Low	Close	Adj Close	Volume
28/10/2020	416.48	418.6	406	406.02	406.02	25451400
29/10/2020	409.96	418.06	406.46	410.83	410.83	22655300
30/10/2020	406.9	407.59	379.11	388.04	388.04	42511300
2/11/2020	394	406.98	392.3	400.51	400.51	29021100
3/11/2020	409.73	427.77	406.69	423.9	423.9	34351700

Table 5: Sample stock market data for TESLA

3.3 Data preparation

The iterative process of data preparation is decisive for the performances of the model. It is also the most time-consuming. There are two separate datasets used in this research that each require data preparation. The entire process of data preparation for both datasets is realized on a Python environment. The pandas' library is used almost exclusively to read, write and structure the data in this research. Data scraped, called or uploaded from comma-separated value (CSV) files are transformed into a pandas' DataFrame, an efficient tabular data structuring frame for working with large datasets. Next, we will discuss each step of the data preparation phase.

3.3.1 Stock market data preparation

On a Python environment, Yfinance is used for extracting stock data for each meme stock. Each of the meme stocks used in this research has its share price information retrieved from Yahoo Finance. In addition to the ready-to-use financial data, two moving averages are added to the dataset that are required as input for time series prediction models (Ko, 2012). According to Droke (2001), 30 days is the ideal period for a moving average regarding short term trading due to its relatively fast signaling of price and volume movements. Therefore, the 30-day moving average of volume and closing share price are added to the stock market data as input for the short-term share price prediction.

30-day volume moving average (VMA30)

$$\text{VMA30} = \frac{\text{Volume } t_{-1} + \text{ Volume } t_{-2} + \cdots + \text{ Volume } t_{-30}}{30}$$

Share price 30-day moving average (SMA30)

$$\text{SMA30} = \frac{\text{share price } t_{-1} + \text{ share price } t_{-2} + \cdots + \text{ share price } t_{-30}}{30}$$

These two trailing moving averages are useful to provide contextual information of the current volume and share price in relation to the averages of the previous 30 days. The share price trailing average is additionally valuable as a baseline input variable to the share price prediction models (Ko, 2012). Both the calculated moving averages are added to the financial data from Yahoo Finance for each meme stock. For the sole purpose of calculating the 30-day moving averages, the last 30 trading days in 2019 are added to the dataset but not used as input to the prediction models.

3.3.2 WallStreetBets data preparation

The WallStreetBets data is retrieved from the Reddit posts directly using the web scraping pushshift.io API on a Python environment. The API scrapes all the submitted data in a certain timeframe on Reddit's WallStreetBets. The metadata that can be saved from scraping all the posts is much more extensive than what is needed for this research. Each post has 81 metadata retrievable objects. Only 4 of these objects are deemed valuable for this research which are the timestamp, title, score and number of comments. Our data from 2020 to 2022 consists of more than a million submitted posts. The data is cleaned and structured with the pandas' library on a Python environment. Any data points that are void or have null values are removed. The resulting dataset is then saved with UTF-8 encoding, a type of textual encoding that is used and preferred because it is possible to read and display emojis. When the data is used for training the machine learning models, it is randomly shuffled and split up into test and training data. In our case, a copy of the dataset is saved after the random shuffle for the purpose of reproduction and is used for the next preparation steps. The WallStreetBets data is prepared for sentiment analysis first and the prediction models later.

Two sentiment analysis models are used in this research. Both models are compared by their performances, as only one model will be used to analyze our data. Before the sentiment analysis can be performed, some Python coding is performed to prepare the dataset. The WallStreetBets posts' sentences are parsed, meaning split up into words, and then lower-cased. Next, stop words, punctuations and pronouns are removed from the parsed sentences. These have no distinctive value to the sentence and can thus be removed for faster analysis. Lastly, the verbs are lemmatized or adjusted to their normal form. An example of a raw sentence into the prepared data is as follows:

raw: GME has been outperforming the market for weeks now, GME to the moon!!
prepared: GME be outperforming market weeks GME moon

The first of the sentiment analysis methods is the SpaCy TextBlob. TextBlob is originally trained on an English dictionary, where a support vector machine is used as a method to classify words into either positive or negative sentiment. It is a simple method and is not custom-made for usage on the WallStreetBets dataset. A training set is therefore not needed in this method. It analyzes the parsed data and consequently categorizes the words as positive or negative. While it is a fast an efficient method, the lingo of the WallStreetBets community poses some difficulties and requires a more custom trained model.

The second method is a custom-made SVM sentiment analysis from the sklearn library. For this method, training data must be provided to the sentiment analysis model. That means manually coding the sentences into positive and negative sentiment. For the SVM model 5000 sentences were scanned and only the obvious negative and positive sentences were marked by either a 0 or a 1, negative and positive (Cieliebak, 2022). The obvious polar posts are taken as a training set because a considerable number of posts have a neutral or unclear sentiment. A positive example is: 'GME has been outperforming the market for weeks now, GME to the moon!!', 1. In total, 500 sentences are classified and then provided as training and test data to the model. It is crucial for the performance of the model to accurately annotate sentences in categories (Mohammad, 2016).

3.4 Modelling sentiment analysis

Both the TextBlob and the SVM sentiment analysis models are constructed on a Python environment. TextBlob reads the prepared WallStreetBets dataset and classifies the sentiment of each post. The result is a Python DataFrame with the date, title and sentiment of each of the post. The sentiment is categorized between -1 and 1, negative to positive, respectively. Additionally, the TextBlob model returns the words from the analyzed post that influences the sentiment categorization. A positive sentiment example could be the word 'amazing' in a post. Such words are useful to identify which ones the model categorize. However, they are irrelevant for further use in the prediction models.

After training the support vector machine (SVM) model, it can be used to predict sentiment on data structured similarly to the training data. The result of predicting a dataset with the SVM model is a pandas' DataFrame in which each post is returned with its predicted sentiment, which is either a 0 or a 1 categorizing them in negative or positive respectively.

3.5 Evaluation sentiment analysis

The best performing method for sentiment analysis is further used in this research. Concluding which of the two models is the best requires evaluation measures. A confusion matrix is a great tool for measuring the model's performance. In the evaluation phase we will focus on the recall and precision of the model as well as the overall F-score.

3.6 Data preparation multiple linear regression (MLR) and Gated Recurrent Unit (GRU)

When the WallStreetBets data has been analyzed and categorized, it can be used to try and find correlations with the stock market data. The DataFrame of the WallStreetBets dataset has a timestamp of each post and could be linked to the stock market date. The first step is separating the posts into five datasets, one for each meme stock. The categorization of each meme stock is done by their textual mentions in the WallStreetBets posts. If a meme stock is mentioned, it will be tagged using the SpaCy rule-based matcher function, which can categorize text by using specific linguistic rules. In an example for GameStop, the different ways of mentioning the stock are implemented as a rule. For example, $GME, GameStop and gme are tagged as GME. The result of the matcher's categorization is the five separate datasets split by meme stock tag. Sentiment is only categorized when the data is separated by meme stock, which reduces the size of the data that needs to be analyzed significantly.

The next step in organizing the data is to create a dictionary in the Python environment. A dictionary has keys and corresponding values that cannot be altered within the dictionary. In the dictionary made for WallStreetBets, the keys are the dates of the posts, and the values are a list of the sentiment values of the posts on that day. A dictionary is created for each of the five meme stocks with the separate datasets.

Each of the dictionaries is then linked back to the financial stock data. The result is a DataFrame that has the common dates as an index. Those dates are the trading days of the stocks that also have at least one mention in a post on the subreddit. We have chosen to use the average of the sentiment score and comments of each post instead of displaying all the posts independently. That is mainly due to the otherwise additional complexity of the input data for the prediction models. The columns on the merged DataFrame display all the relevant financial stock data as well as the gathered sentiment data. Table 6 is a sample of an unsorted DataFrame of the merged data that can be used for multiple linear regression or GRU.

The merged DataFrame is normalized for the GRU model and the linear regression evaluation metrics. Each variable's value range is fitted between zero and one. The absolute highest closing share price has a value of one, and vice versa for the lowest share price. By normalizing the date, the evaluation metrics can be easily compared to each other. The reversal of normalization is used for the GRU model output, which results in a normal range of share price prediction.

Date	Sentiment	Score	Comments	Posts	Close	SMA30	VMA30
2021-03-05	0.187500	240	58	16	406.02	782.41	31077490
2021-03-11	1.000000	1832	82	5	410.83	753.40	34559630
2021-03-04	0.636364	63	32	11	388.04	790.64	28781640
2021-04-22	1.000000	193	59	5	400.51	685.77	34132100
2021-01-29	0.555556	928	21	63	423.9	768.28	44655740

Table 6: Merged data from TSLA

3.7 Modelling multiple linear regression

Each of the meme stocks analyzed in this research has its own merged dataset. The 'Close' value of the stock is used as the dependent variable. That is what the MLR model will try to predict. As the independent variables we have chosen the average sentiment of all the posts on the certain day, the number of posts, the score and their comments, along with the 30-day moving averages of the share price and volume. A model is constructed for each of the five meme stocks. The MLR model is trained and constructed using the sklearn library on Python, where the test and training sets are split and then trained.

3.8 Modelling Gated Recurrent Unit

The GRU model is constructed using the Tensorflow library. Here, we use the normalized version of the dataset that is used for the multiple linear regression. A multi-variate forecasting model is trained for each of the meme stocks and is used for predicting the next day's stock market closing price. The Tensorflow library together with the Keras API offers customization in building the GRU model. The options for activation and recurrent activation are the hyperbolic tangent and the sigmoid function, respectively. Those are the default options and behave similarly to each other. When the model trains, it takes the independent variables from the 10 previous days to predict the next day's share price. The neural network has three layers and is iteratively trained 100 times while minimizing the mean squared error. We opted to train the GRU models with the same complexity for each of the meme stocks. The same number of layers of the GRU neural network are added to each model.

3.9 Evaluation MLR and GRU

MLR models are evaluated by several evaluation methods. Every model is evaluated with the R-squared value, mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE), as well as the coefficients for each of the independent variables. Each meme stock model's usability and effectivity are analyzed by comparison. The evaluation methods for MLR are discussed in the literature review. The GRU models are evaluated by the MSE and RMSE error measures. The usability is considered by reviewing the GRU model's graphed prediction share price compared to the actual share price. Due to the nonlinear nature of the GRU model, the R-squared value is not a useful measure (Spiess & Neumeyer, 2010).

4. Results

4.1 Business understanding

Stock	Date	Peak share price	P/E (peak)	EV/EBITDA (2021)	DCF[4]
TSLA	04/11/2021	$1229.21	202.25	174.4x	$172.09
GME	27/01/2021	$347.51	Negative	Negative	$37.04
BB	27/01/2021	$25.01	Negative	95.2x	$3.30
AMC	27/06/2021	$62.55	Negative	Negative	$4.81
PLTR	27/01/2021	$39.00	Negative	Negative	$5.75

Table 7: Technical analysis meme stocks

Table 7 depicts some of the meme stocks financial data and valuation measures. Three of the five meme stocks traded at an all-time high on the 27th of January 2021. GME, BB, AMC and PLTR all had negative P/E ratios at their peak share price which means that out of the five meme stocks only TSLA reported positive earnings in 2021. At Tesla's peak share price in November 2021 its P/E ratio was 202.25. While that is an extremely high figure, TSLA once traded at an enormous P/E ratio of more than 1200. When compared to other electric vehicle manufacturers, TSLA remains very highly valued in their industry.

Looking further at the EV/EBITDA ratio, most meme stocks do not present healthy financial figures. GME, AMC and BB do not have a positive EV/EBITDA ratio, which makes the ratio not useful. A negative ratio signifies that they have a negative cash flow or EBITDA, which is concerning. The ratios for TSLA and BB are positive but are immensely high. The average S&P500 EV/EBTIDA ratio is 17.12 (Enterprise Multiple by Sector/Industry 1995 – 2022, 2022). A high ratio means that the equity value, consisting partly of market capitalization, is overvalued. Tesla is especially overvalued when compared to the average ratio.

Discounted cash flow (DCF) is a method for valuating a firm's value. A fair share price can be projected by dividing the valuation of the firm by the outstanding shares. In this case, the DCF figures divided by their outstanding shares are gathered from 'Finbox.com'. These values should represent the honest share price for the meme stocks at their peaks. All the meme stocks' DCF values are considerably lower than their peak share price. In proportion, the most honest valued meme stock is Palantir Technologies, which has a peak share price six times greater than its DCF value. AMC entertainment is the worst performing stock, with its peak share price 13 larger than its DCF value.

[4] (*The Complete Toolbox for Investors*, 2022)

Table 7 is a great demonstration of the extent of the meme stocks overvaluation. All the meme stocks show clear signs of overvaluation when examining the different financial ratios. Tesla is the only company that has a positive price to earnings ratio and has an expected price that is somewhat reasonable.

4.2 Data preparation

In this section, we will go over the data preparation for the stock data and the sentiment analysis models.

4.2.1 Data preparation stock market

Date	Open	High	Low	Close	Volume	SMA30	SVMA30
11/08/2020	279.2	284	273	274.878	43129000	289.3012	74793483
12/08/2020	294	317	287	310.952	109147000	292.4675	75611967
13/08/2020	322.2	330.236	313.452	324.2	102126500	295.81	76795033
14/08/2020	332.998	333.76	325.328	330.142	62888000	298.757	76016283
17/08/2020	335.4	369.172	334.566	367.128	101211500	301.8507	75961683

Table 8: Sample stock market data for Tesla including SMA30 and SVMA30

Table 8 illustrates the imported stock data for Tesla from Yfinance. The trailing 30-day moving average of share prices and the 30-day moving average for volume are calculated in Python and added to the dataset. This dataset is later merged with the WallStreetBets (WSB) data for the prediction models. Due to the calculations of the moving averages, the range of the stock data is broader than the WallStreetBets data. The first 30 rows are solely needed for the calculation of the moving averages and are subsequently dropped from the merged dataset ranging from 2020 to 2022.

4.2.2 Data preparation WallStreetBets

In this research, we gather our data by scraping it off Reddit using the API from pushshift.io. The extraction script is written on a Python environment and captures data from WallStreetBets between 2020 and 2022. The dataset is exported in a textual format and saved. The textual data is then prepared in a Python environment and converted to a pandas' DataFrame. In total 1,456,167 posts were scraped off WallStreetBets including their title, score, number of comments and date. Not all these posts are relevant to our research. The interesting data points are the ones that mention at least one of the five meme stocks. It is also helpful to slim down the large dataset before it is categorized by the sentiment analyzer. To that effect, the SpaCy matcher is used to separate the dataset into smaller ones by mention of meme stock. Any form of text that represents a meme stock receives a tag by the matcher. In the WallStreetBets dataset, more than 20% of the posts has at least one mention of a meme stock.

GME Pattern
[{"LOWER": "gamestop"}]
[{"LOWER": "gme"}],
[{"IS_PUNCT": True}, {"LOWER": "gme"}]]

Table 9: SpaCy matcher pattern GameStop

A sample rule set for GameStop is shown in Table 9. The matcher uses a list of patterns or rules. Each rule is a list of dictionaries. The combination of ‘is_punct equals true’ and the ‘lower equals true‘ results in any form of punctuality in combination with the word gme to be picked up by the matcher. The common form of ticker notation ‘$GME’ would be matched by that rule.

Next, the data is prepared for annotating the WallStreetBets data into polar categories. Manual categorizations are necessary for the model to determine the sentiment polarity. The annotations are added to the dataset, which is then used as training data for the sentiment analysis models. Before annotating, the practical guide to annotating sentiment by Mohammad (2016) was consulted. Using this guide, we decided that a small dataset of 500 manually annotated titles were sufficient. In the training data, we used the obvious positive and negative titles so that the model could accurately learn to distinguish both. Table 10 shows a sample of the manually annotated data that is used as input for the support vector machine (SVM) sentiment analysis model. The other sentiment analysis model of TextBlob does not require an annotated dataset, as it has already been trained on English dictionaries.

Title	Target
Smells like lawsuit to me	0
WHAT DO WE LIKE??? 👐 💎	1
AMC IS TANKING	0
I’m In You Monkey Futher Muckers 🐒 🚀	1
Still holding strong and yes I do own stock but that’s on my other account my Robinhood is pure options	1
I'm so proud of you guys holding the line. They tried so hard end of day to drive down the price below $320 - look at this volume on the 5 min chart.	1

Table 10: Sample annotations WallStreetBets

4.3 Modelling WallStreetBets sentiment analysis

The modelling of the sentiment analyzers is done on a Python environment. The two models used are retrieved from the libraries TextBlob and sklearn. The first is a simple and already pretrained TextBlob English sentiment analyzer that is called from SpaCy. It only requires the input that needs to be categorized. The second sentiment analyzer is the sklearn custom analyzer that uses the SVM algorithm and requires manually categorized data. Here the data is split in training and test data with a 20% to 80% Pareto Principle split respectively (Dunford et. al.,2014). This is to prevent the model from overfitting. The result from both sentiment analysis models is an effective way to categorize the sentiment of all the posts in the dataset.

4.4 Evaluation WallStreetBets sentiment analysis

In this section, both sentiment models will be evaluated with confusion matrices and their recall precision and F-score. The best performing model will be used to predict the sentiment of all the WallStreetBets posts data.

4.4.1 TextBlob sentiment analysis

		Actual	
		Positive	Negative
Predicted	Positive	55	41
	Negative	148	256

TextBlob	Value
Recall	0.271
Precision	0.573
F-score	0.422

Table 11: Confusion matrix and evaluation measures TextBlob

The model of the TextBlob simple English sentiment analysis does not perform very well. The overall F-score of the model is 0.422, meaning that the model only predicts 42% of the posts correctly. While the precision is not remarkable at 0.573, it is the recall especially that is terrible. This is mainly due to the model's incapacity of recognizing true positives while at the same time predicting too many false negatives. The model does perform accurately enough to be used as a reliable sentiment analyzer for the WallStreetBets data. Next, we will compare this performance with the manually trained SVM sentiment analyzer.

4.4.2 Sklearn SVM sentiment analysis

		Actual	
		Positive	Negative
Predicted	Positive	127	37
	Negative	76	260

SVM	Value
Recall	0.626
Precision	0.777
F-score	0.700

Table 12: Confusion matrix and evaluation measures sklearn

The SVM model is trained on manually annotated data and has noticeably better performances than the previous model. It has similar values for true negatives but has considerably fewer false negatives in comparison to the TextBlob model. Unlike the TextBlob model, the SVM model does well to recognize the true positives. This model predicts similar numbers of false positives, but these errors are largely negated by the larger amount of correctly predicted positives. The model's precision and recall have values that are close to each other, which is desirable. This model has an overall score of 70% to categorize the sentiment of the WallStreetBets posts. When compared to the previous model, it is obvious that we will choose this model over the TextBlob to further analyze the correlation between the sentiment and stock market performances.

4.5 Data preparation multiple linear regression and GRU

Each meme stock has its own multiple linear regression analysis and recurrent neural network model using GRU. The models require both dependent and independent variables. The dependent variable is the closing share price of the meme stock and requires no additional preparation. The independent variables are the mean sentiment, number of posts, mean number of comments, mean score of the WallStreetBets title data merged with the SMA30 and VMA30 that are calculated from the stock data. These independent variables are then added to a pandas' DataFrame, with their dates as index. An additional preparation is necessary for the errors evaluations of the linear regression model and the GRU model, which require normalization. Each value within a column or variable is modified to a relative value between 0 and 1. As an example, the highest stock market close in the dataset has an adjusted value of 1, the lowest value 0. The normalization is necessary to objectively compare the models to each other.

4.6 Modelling multiple linear regression

The multiple linear regression is modelled on a Python environment with the sklearn library. The model requires dependent and independent variables. The input is the merged DataFrame per meme stock that we have discussed in the data preparation phase. A separate model is made for each of the meme stocks. The predicted share price values are visualized by a graph together with the actual values. The evaluation metrics for the linear regression are also calculated during the modelling phase.

4.7 Evaluation multiple linear regression

Next, we will discuss every meme stock's multiple linear regression model. Table 13 displays an overview of the most important evaluation metrics for each model. The R-squared is displayed by a percentage.

Evaluation	TSLA	GME	BB	AMC	PLTR
R-squared	94.90%	88.53%	60.53%	84.29%	46.63%
RMSE	0.064	0.096	0.111	0.115	0.113
MAE	0.043	0.057	0.058	0.080	0.089
Number of predictions	148	123	84	107	70

Table 13: Overview of multiple linear regression models

4.7.1 MLR Tesla

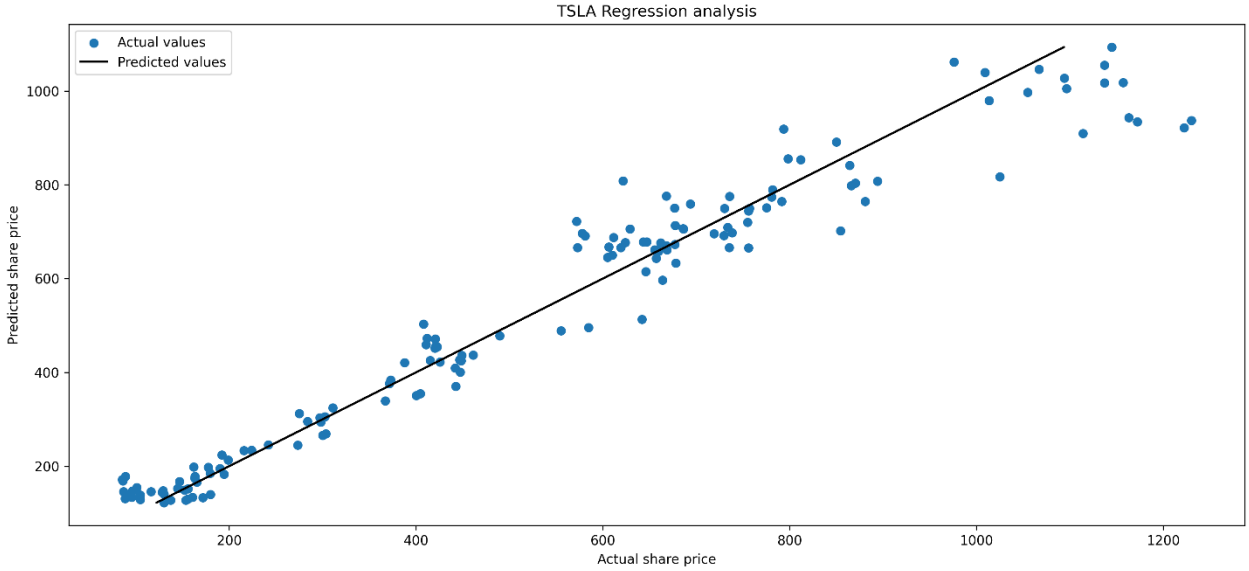


Figure 11: TSLA Regression analysis

The first meme stock model that is evaluated is Tesla. The multiple linear regression model predicted 148 share prices randomly selected between the timeframe of 2020-2022.

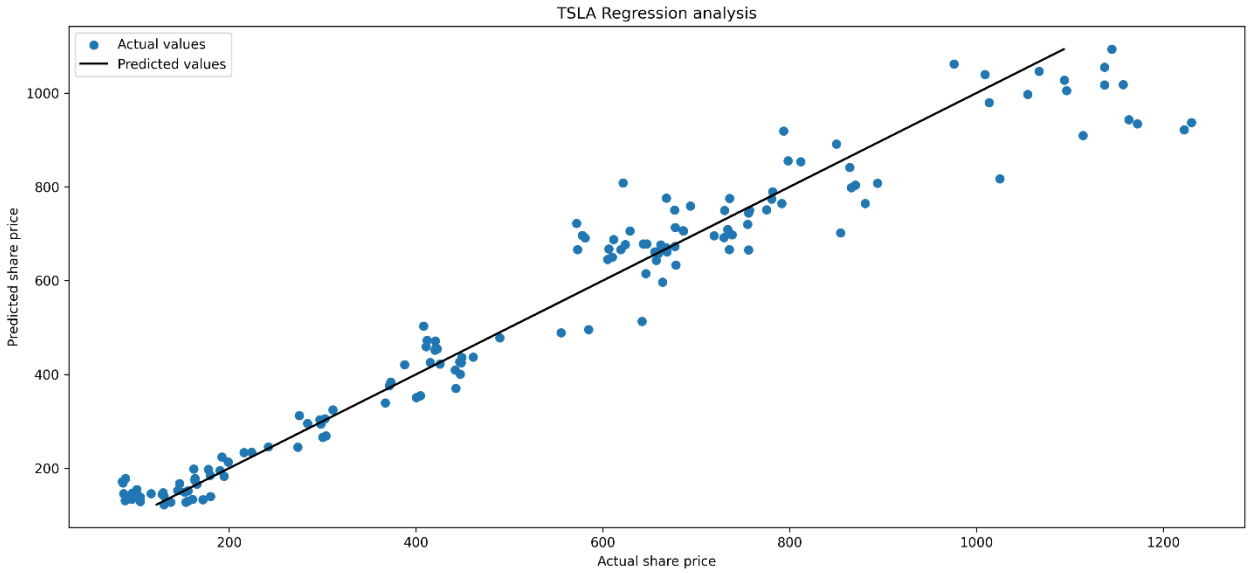


Figure 11 shows how close the model's predictions for Tesla's share prices are compared to their actual values. The scatter plot portrays the actual values and the best fit line of the regression model. If the prediction model would be 100% correct, all the dots would be on the regression line of the predicted

values.

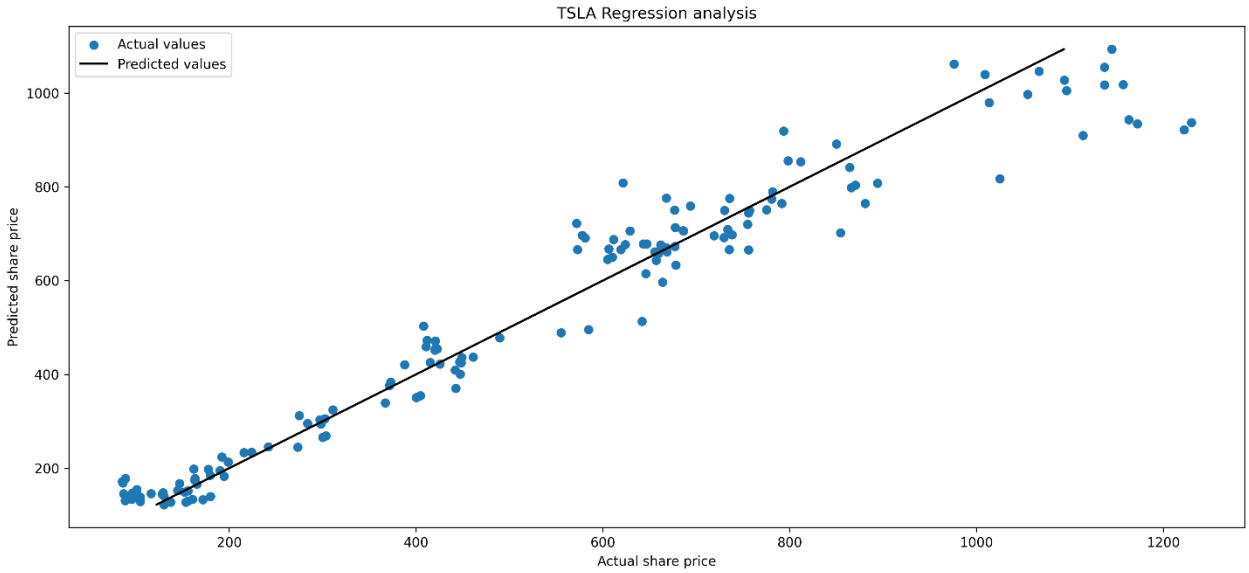


Figure 11 is an ideal representation of the effectiveness of the model's prediction strength. The R-squared value of this model is 94.90%, which is a great score. The R-squared value is a measure for how well the independent variables explain the variability of the dependent variable. In this case, Tesla's share price can be 95% explained by the independent variables, suggesting a quite robust model.

TSLA	R-squared	MAE	MSE	RMSE
Values	94.90	0.044	0.004	0.064

Table 14: Regression analysis TSLA

Looking further at the measures for evaluating the errors of the model in Table 14, the Tesla model has low error values. Compared to the other models, it consistently scores the best. It is important to note that the DataFrame has been normalized for the error evaluations, so that their scores can be objectively compared. The model's MAE, MSE and RMSE have the lowest values relative to the other models, meaning that it in comparison it is highly effective in predicting share prices when given the independent variables.

TSLA	Intercept	sentiment	score	comments	posts	SMA30	VMA30
Coefficients	54.449	37.755	-0.097	0.802	0.136	0.946	-5.467

Table 15: Regression coefficients TSLA

Next, we will review the coefficients of the Tesla model in Table 15. The coefficient of sentiment as an independent variable stands out. Its high value suggests that sentiment has a lot of influence in the model's share price prediction. One unit of sentiment, whose values are always between zero and one, will result in a 37.755 difference in the predicted share price. Another coefficient that stands out is the volumetric moving average of the share price. The VMA30 has a negative coefficient of 5.467, which indicates that it has a strong negative impact on the share price prediction.

4.7.2 MLR GameStop

Figure 12: GME Regression analysis

GameStop's (GME) multiple linear regression model clearly has some more issues to predict share prices when compared to Tesla. The high values in share prices are particularly hard to predict, as seen by the outliers on the right side of the x-axis. These outliers correspond to the period in January 2020 when GME's share price closed at more than 300 during the short squeeze. When comparing Figures 11 and 12, it suggests that a spread-out scatter plot corresponds to a worse R-squared value. Another noticeable difference between the two figures is that GME's scatter plot has a high number of share price values at the higher and lower end but lacks data points in between.

GME	R-squared	MAE	MSE	RMSE
Values	88.53	0.057	0.009	0.096

Table 16: Regression evaluation GME

Still, the model's effectiveness is relatively strong, as the R-squared value of the GME model is 88.53. Despite the stock's volatility in share prices, the model still performs well with its 123 predictions. Examining further, the RMSE and MAE are only fractionally higher than the TSLA's high performing model. RMSE especially punishes outliers, which GME clearly has more of.

GME	Intercept	sentiment	score	comments	posts	SMA30	VMA30
Coefficients	-0.677	2.975	-0.168	-0.006	0.005	0.976	4.729

Table 17: Regression coefficients GME

Analyzing the coefficients of the GME model, the sentiment variable is a smaller predictor than in the model for Tesla. However, the volumetric moving average is similarly impactful in the prediction of the regression model but has an opposite sign. Another notable pattern is the steady coefficient of the simple moving average of the stock price in the last 30 days. Both TSLA and GME's models have similar values of just under 1.

4.7.3 MLR BlackBerry

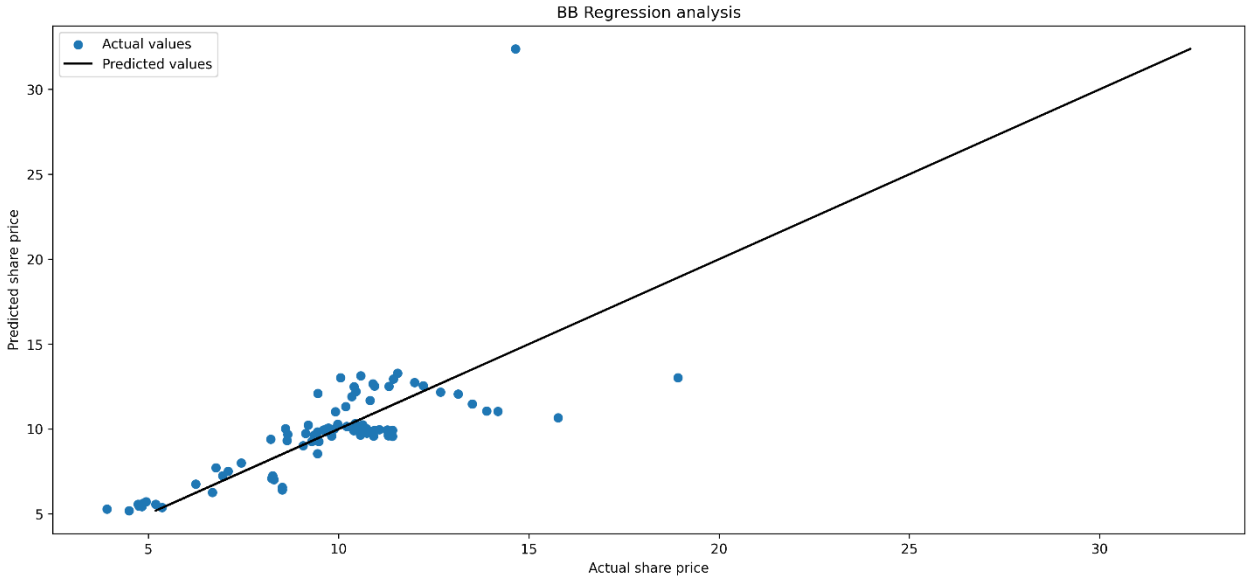


Figure 13: BB Regression analysis

BlackBerry's model does not perform effectively. The model predicted 84 share prices, which is 30% of the dataset size. This is considerably less than the previous two models. That is due to the limiting factor of being mentioned in the WallStreetBets posts. Clearly, BlackBerry is less mentioned than the previous

two meme stocks. When looking at Figure 13, again, the model struggles with the outliers that are the meme stocks price peaks. The scattered actual values of the lower left corner seem quite closely positioned to the regression line. The R-squared value however suggests that the model did not perform effectively.

BB	R-squared	MAE	MSE	RMSE
Values	60.63	0.058	0.012	0.111

Table 18: Regression evaluation BB

The R-squared value of BB's model dips considerably compared to the previous two models. In this case, the independent variables can only explain 60.63% of the share prices. The model is clearly weaker than the previous two, which can also be identified by the error evaluations. The RMSE jumps up to 0.111 resulting in almost double the error than the Tesla model, while the MAE retains similar values when compared to the GME model.

BB	Intercept	sentiment	score	comments	posts	SMA30	VMA30
Coefficients	1.699	0.171	-0.002	-0.001	0.003	0.771	1.571

Table 19: Regression coefficients BB

The coefficients of the BB model have smaller values. The independent variables from the WallStreetBets data do not contribute much to the share price prediction. The share price and volumetric moving averages have the most impact.

4.7.4 MLR AMC Entertainment

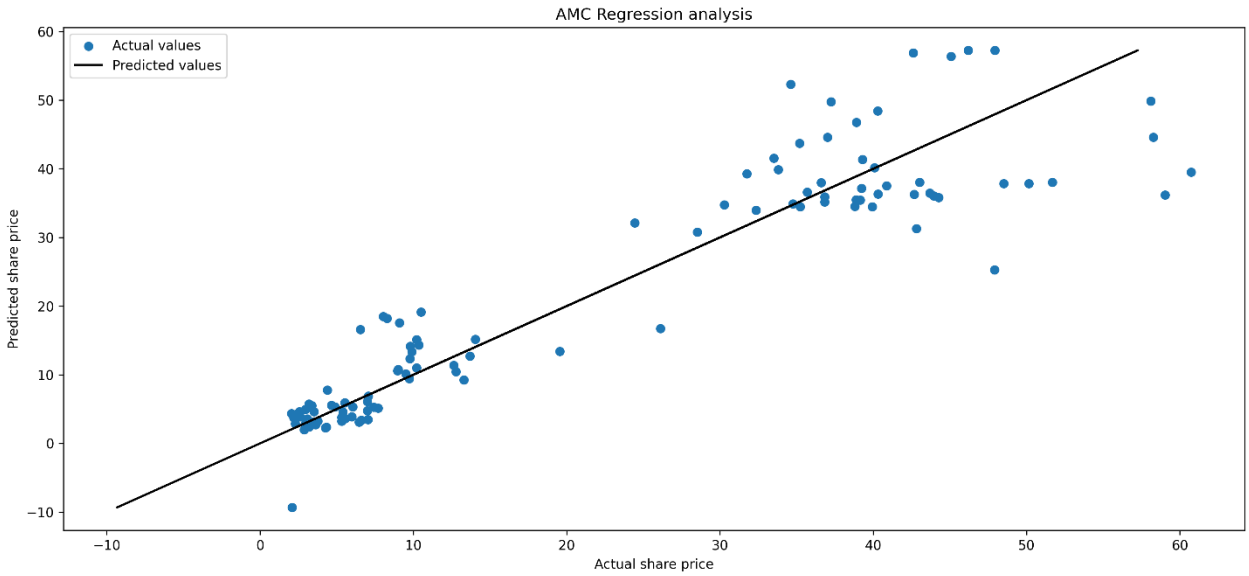


Figure 14: AMC Regression analysis

AMC's multiple linear regression model performs well in predicting share prices. Most actual values sit along the regression line of the predicted values in Figure 14. When compared to BB's regression analysis, AMC's values are spread more proportionally along the predicted values line, which suggests a better performance. Curiously, one prediction turned out to have a negative value. This is clearly an error made by the model, which will be penalized by the error evaluations.

AMC	R-squared	MAE	MSE	RMSE
Values	84.29	0.08	0.013	0.115

Table 20: Regression evaluation AMC

As suggested by Figure 14, the model has strong predictive power, indicated by the considerably high R-squared figure of 84.29. However, the error evaluation values have similar or higher values than the lower performing BB model. So, while the independent variables effectively explain the closing share price, the model produces errors that are similar to the BB model. As a whole, this model does not produce the same performance levels as in the GME or TSLA models.

AMC	Intercept	sentiment	score	comments	posts	SMA30	VMA30
Coefficients	-1.607	1.743	-0.133	0.043	0	0.876	5.34

Table 21: Regression coefficients AMC

Notable coefficients in the AMC regression model are the sentiment, SMA30 and VMA30. The VMA30 has the strongest positive impact on the predicted share price. The sentiment has a rather small positive effect, while the coefficient for the share price moving average behaves similarly to the SMA30 coefficients in the previous models.

4.7.5 MLR Palantir Technologies

Figure 15: PLTR Regression analysis

In Figure 15, it is clear that PLTR's model has had major difficulties in constructing an accurate share price prediction. PLTR only started trading in the last quarter of 2020, which has resulted in a smaller dataset to train on. 70 predictions were made by the model. Compared to all the other models, this one has the least dense grouping of actual values close to the predicted values line. Additionally, there are multiple outliers situated on the right top side, one of which is extremely far from the predicted values.

PLTR	R-squared	MAE	MSE	RMSE
Values	46.63	0.089	0.013	0.113

Table 22: Regression analysis PLTR

Table 22 shows the evaluation values of PLTR's model. The R-squared value is the lowest of all the models. Similarly, this model has the highest mean absolute error, while the RMSE is nearly identical to the models of AMC and BB. All in all, this model is clearly not as effective as the previous ones.

PLTR	Intercept	sentiment	score	comments	posts	SMA30	VMA30
Coefficients	10.205	0.007	0.068	-0.063	0.022	0.604	-1.300

Table 23: Regression coefficients PLTR

The multiple linear regression model for PLTR did not identify any data from WallStreetBets as impactful. All the coefficients apart from the moving averages are small and do not influence the share price predictions much.

4.8 Modelling GRU

The recurrent neural networks, specifically the GRU, is modeled on Python using the Tensorflow library together with Keras, an API for building deep learning models. Keras allows for the customization of recurrent neural networks. In our case, we have built a GRU neural network with three layers and 100 iterations. The GRU is trained with a loss function that focuses on mean squared error. Other larger sized neural networks have been informally experienced with, but the added complexity resulted in declining results. The normalized DataFrame is used as input for the model, while the output is transformed back to normal size values. Those normal values aid to better understand the predictions of the GRU model. An important note for the modelling with GRU is that the randomization of the data is not possible due to the sequential nature of recurrent neural networks. The model is trained on the first 80% of the dataset, meaning the dataset has a timeframe from 2020 to just after Q3 2021. The testing happens with the last 20% in which the model predicts the share prices. We can compare those to the actual share price values of the meme stocks in the second half of 2021. A GRU neural network is made for each of the meme stocks along with evaluation metrics.

4.9 Evaluation GRU

In this section, we will evaluate the GRU models. Table 24 provides an overview of all the models for each meme stock, which we will cover next.

Evaluation	TSLA	GME	BB	AMC	PLTR
RMSE	0.116	0.070	0.075	0.232	0.130
MSE	0.013	0.005	0.006	0.054	0.017

Table 24: Overview GRU models

4.9.1 GRU Tesla

As with the multiple linear regression model, the Tesla GRU model visually performs the best out of the five meme stock prediction models. However, TSLA's GRU model is not as effective as its regression counterpart. The GRU model clearly has difficulties with predicting the stock market values, as is seen in Figure 16. While the model does show similarities in large trends, it seems to lag in its predictions. The RMSE and MSE evaluations for the TSLA GRU model are acceptable and can be compared to the regression models. Compared to the four other GRU models, Tesla's model has mid-range error values.

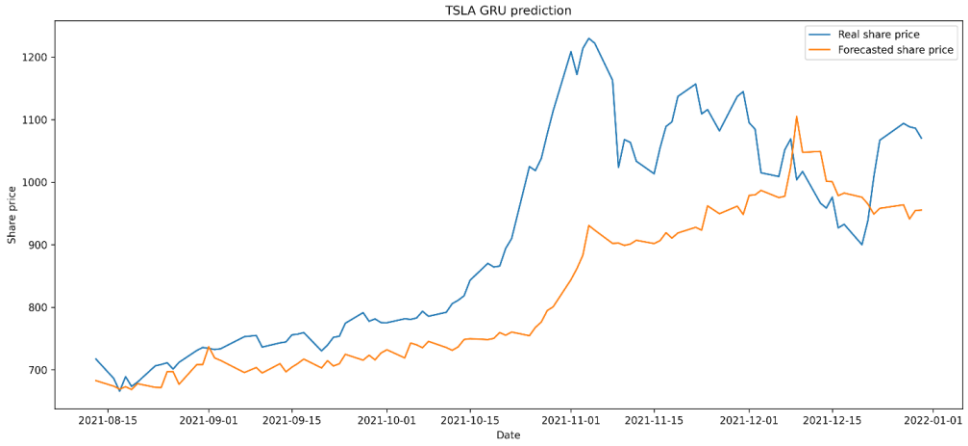


Figure 16: TSLA GRU prediction

4.9.2 GRU GameStop

GameStop's GRU model is not a good predictor for the share prices. Visually, it can be clearly seen in Figure 17 that the GRU model fails to accurately predict the volatility. There is also a notable delayed drop in the predicted share price, a couple of days into a downtrend. While this GRU model has the lowest error margins, it is definitely not a usable model when comparing the actual values to the predicted values.

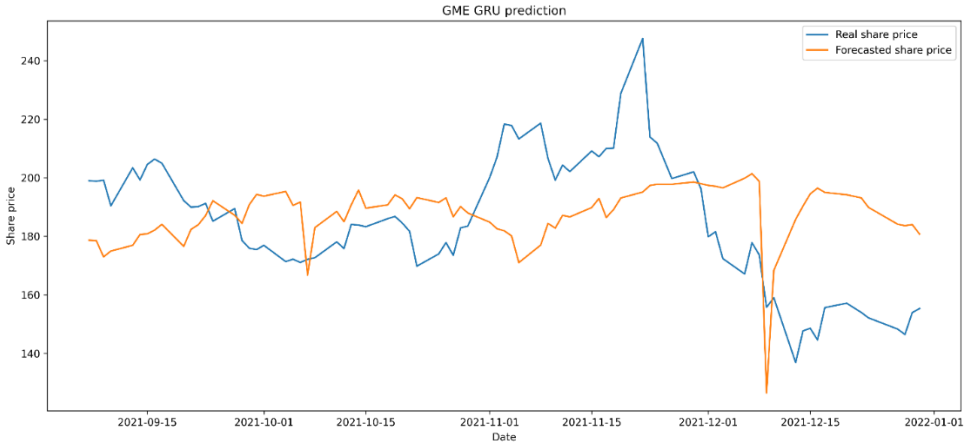


Figure 17: GME GRU prediction

4.9.3 GRU BlackBerry

BlackBerry's GRU model is a horrible looking model. The model does a terrible job at predicting short-term share prices, as can be seen by the Figure 18. For the most part of the graph, the forecasted values seem to be a mirror image of the actual values. When looking at the error measures from Table 24 however, the BB model does show some strength. It has the second lowest MSE and RMSE values of the five meme stock models at 0.006 and 0.075 respectively.

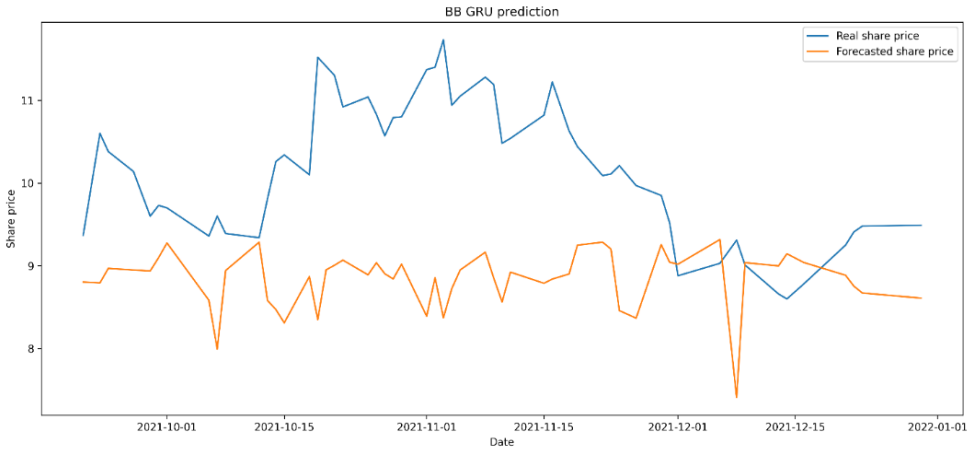


Figure 18: BB GRU prediction

4.9.4 GRU AMC Entertainment

Another model that has failed to deliver a usable share price prediction is the AMC GRU model. Here, the forecasted share price is almost a smooth line that demonstrates similarities to a moving average. Throughout the whole prediction period of two months, not once does the predicted value cross the actual share price values. The model consistently overestimates the share price, and as a result, it has the highest error values of the five models with an RMSE of 0.054.

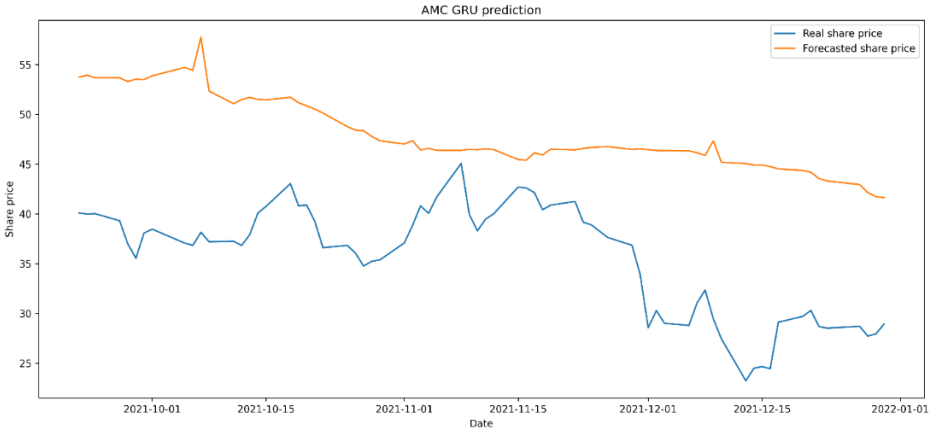


Figure 19: AMC GRU prediction

4.9.5 GRU Palantir Technologies

The last GRU model is Palantir's. Here, the model visually performs better than the previous two. Up until the 2021-12-01, it seems like the model in Figure 20 does follow the real share price very closely. For a two-month period, the model seems at least somewhat usable. Later however, the predictions by the model are far off the actual share prices. The error measure is the second worst of the five models, with an RMSE of 0.130.

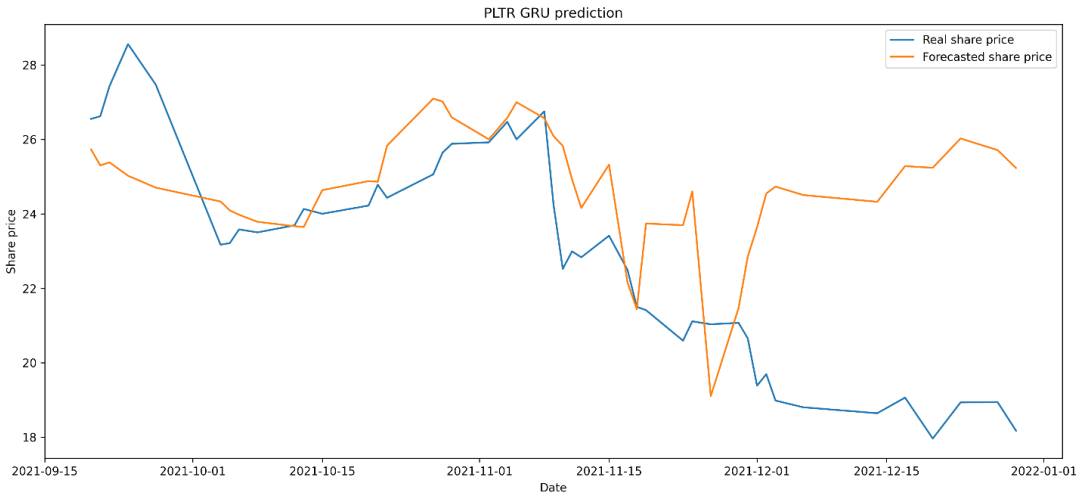


Figure 20: PLTR GRU prediction

5. Discussion

In this section, we will interpret the results of the research examined in the previous section thoroughly. We will start by reviewing our data gathering methods and later discuss the performances of the machine learning models used.

5.1 Data gathering

In total, we have gathered 1,456,167 posts from the subreddit of WallStreetBets. 296,443 (20.36%) of those have mentions of at least one of the five meme stocks that are researched. That is a surprisingly high number. The high percentage of the mentions is indicative of the relevancy of the stocks that we have chosen for this research. The WallStreetBets dataset that is used for the five different linear regression and GRU models is large and can be automatically retrieved by a script. Pushshift.io proved to be an effective API for the collection of data from Reddit. Supposing that the models are effective, it would be entirely possible to collect and update data instantly using the API. The other dataset that was used for modelling share price predictions is the stock market data retrieved by Yfinance. This is an effective library for collecting all the necessary stock data. It works instantly and was particularly clean.

Some media reports suggest that bots are used to influence sentiment of certain stocks on Reddit (Gandel, 2021). Bots are non-human users that automatically post and are usually coordinated. This could potentially be a problem in our research case. However, we have used such an extensive dataset that it should not skew our results. In conclusion, our data gathering method was justified and resulted in clean datasets.

5.2 Sentiment analysis

We have modeled two different sentiment analysis models for this research. After testing their performances, we chose the better performing support vector machine (SVM) model by sklearn as the sentiment analysis model of choice. This model can predict 70.00% of our WallStreetBets data correctly by sentiment. However, this figure overestimates the model's performance due to the manual annotations that it is tested on. The posts that were too not obviously positive or negative were left out of the dataset. As a result, the sentiment analysis model learned from a cleaned-out dataset without room for interpretation before it was released to be used for the 296,443 posts. The real performance of the sentiment model is not as reliable in reality.

Aside from this limitation, there is also a point to be made about vocabulary used by the WallStreetBets community. The jokes, memes and sarcasm used in the WallStreetBets posts are omnipresent. It is possible that some of the posts that would seem positive to anyone that does not frequently visit the WallStreetBets subreddit is in fact negative. It is evident that this complexity of sarcasm and meme culture

is very hard to recognize by a sentiment analysis model. The sarcasm is one of the difficulties that are mentioned in the practical guide to annotating sentiment by Mohammad (2016). We made the choice to take an aggregate of each day on the WallStreetBets posts. In this manner, the overall sentiment of the posts for each day per meme stock should be more accurate than single posts. These overall values are then taken as an independent variable for the prediction models in the hope that the average sentiment would be more accurately portrayed.

In our sentiment analysis model of choice, we opted for the SVM algorithm in the sklearn library. This decision was influenced by reviewing the literature. Additionally, we have used SpaCy's vectorizer and not the available TF-IDF, which could have enhanced the sentiment analysis model slightly. Many factors could enhance the performances of sentiment analysis models. However, it will remain extremely hard to accurately categorize sentiment on a forum of a vibrant, young online community such as WallStreetBets.

5.3 Multiple Linear Regression

The results from the multiple linear regression analysis are largely strong. Tesla, GameStop and AMC Entertainment models have R-squared values that exceed 80% and have low mean squared errors at the same time. The independent variables overall explain the share price of the meme stocks remarkably well, despite their unpredictable and volatile nature. Examining the independent variables, most of the WallStreetBets data coefficients of the regression models do not have a significant impact on the predictions of the share price. Average sentiment is the only independent WallStreetBets variable where the coefficient is impactful, notably in the best performing regression model for Tesla. This result is an indication that average sentiment has the potential to be significant in the prediction of share prices. The 30-day moving averages of volume and share price are the independent variables that have the most significant impact to the regression models. The SMA30 coefficient is a stable predictor as its values are always between 0.64 and 0.97 while the VMA30 is more variable in size and direction.

The regression models are generally successful in predicting share prices. Most of the time, the independent variables as input of the regression model result in fairly accurate share prices. Times when the regression models' share price predictions are off, despite their effectiveness, is where actual share prices soar to unprecedented valuations. This is especially true for GameStop and other meme stocks in January 2021.

At this time, the multiple linear regression models do not predict one day ahead of time. Instead, they predict an accurate daily closing price given the independent variables. For the linear regression model to be genuinely predicting a share price for the next day, each of the independent variables would have to be

predicted one day ahead of time. It is possible to forecast each of the independent variables, but this will most likely result in a less effective overall prediction model.

5.4 Gated Recurrent Unit

The Gated Recurrent Unit neural networks are not effective or usable at predicting meme stock share prices. Throughout all the models that have been trained by the same independent variables as the multiple linear regression models, only the Tesla and the BB GRU models seems somewhat effective. The error margins are surprisingly low compared to the seeming lack of usability of the models. Mean squared errors are trained by the model's backpropagation to be as minimal as possible. However, the usability of the model is clearly lacking when examining their plotted predictions in the visualizations.

There could be multiple explanations as to why these neural networks fail to be effective as prediction models. The first reason is that accurately predicting stock prices is near impossible. Especially when the stocks are volatile. Secondly, recurrent neural networks are trained on historical sequential data (Amidi & Standford.Edu, 2019). The period between 2020 and 2022 might not be a suitable period for recurrent neural networks to be learning stock price evolution. Another reason why the performances of the GRU models are lacking is that it is tough to examine the neural network's black box decision-making and subsequently try to direct it.

In future research on this topic, different sized GRU recurrent neural networks could be compared to each other. Different activation functions and iterations can be experienced with. There are endless possibilities when it comes to the customization of neural networks. Nonetheless, predicting volatile meme stock prices will always remain a difficult task.

6. Conclusion

In this research, we tried to answer the research question if multiple linear regression or gated recurrent units can be a reliable model to predict meme stocks share prices by analyzing sentiment on Reddit's WallStreetBets. The rise of small retail investors and their coordinated meme stock investments in the wake of the COVID-19 pandemic sent meme stock valuations through the roof. At the time of writing this research, the topic of meme stocks and WallStreetBets influence has not been thoroughly explored yet. Future research on this topic is essential and should be welcomed to add to the minimal literature that exists today.

We adopted the CRISP-DM standard process for data mining as our methodology. First, we scraped all the posts of the subreddit WallStreetBets in the timeframe of 2020 to 2022. These posts were analyzed and used as input for the sentiment analysis. The posts were divided by their mentions of one of the five meme stocks and categorized into positive or negative sentiment. The data retrieved by WallStreetBets for each meme stock is connected to their stock market data, including the calculated 30-day moving averages for volume and closing price. The independent variables for the prediction models are partly stock market data with 30-day moving averages together with the daily average WallStreetBets sentiment, number of comments and score.

Overall, the multiple linear regression models exhibit some strong results and appear usable. This is especially true for the prediction models of Tesla, GameStop and AMC. The Tesla regression model has the best R-squared value of 94.40% and the lowest RMSE value of 0.064. This suggests that the independent variables, including the WallStreetBets sentiment data, largely explain the closing share price. However, to be an absolute predictor, there is a need for additional next day forecasting of the average sentiment, score and number of comments along with moving averages of volume and closing share price. In that case, a closing share price can be predicted for the next trading day. This has not been included in this research.

Another method that has been used to predict stock prices is the gated recurrent unit recurrent neural network. The results of this deep learning method are mostly poor, although the root mean squared values of errors are respectably low. When examining the predictions for each of the meme stocks, it becomes clear that the models are not reliable as predictors and consequently unusable. There are various reasons for the poor performances of the models. Predicting stock prices is tough, particularly in the specific case of meme stocks that tend to rise astonishingly high in share prices. Additionally, the recurrent neural network models learn from an unprecedented span of share price valuations, which might not serve as an accurate representation of future share price.

Reference list

About Palantir. (2022). Palantir. Retrieved May 2, 2022, from https://www.palantir.com/about/

About Tesla | Tesla. (2022). Tesla. Retrieved May 2, 2022, from https://www.tesla.com/about

Ahmed, S. I. (2022, January 28). *Meme stock hangover: a year after GameStop, traders face gloomier markets*. Reuters. Retrieved March 2, 2022, from https://www.reuters.com/business/finance/meme-stock-hangover-year-after-gamestop-traders-face-gloomier-markets-2022-01-28/

Altig, D., Baker, S. A., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., Davis, S. J., Meyer, B. H., Mihaylov, E., Mizen, P., Parker, N., Renault, T., Smietanka, P., & Thwaites, G. (2020). Economic Uncertainty Before and During the COVID-19 Pandemic. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3638649

Amidi, A. & Standford.Edu. (2019, January 6). *CS 230 - Recurrent Neural Networks Cheatsheet*. Stanford.Edu. Retrieved April 5, 2022, from https://stanford.edu/%7Eshervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

Anderson, K., & Brooks, C. (2006). Decomposing the price-earnings ratio. *Journal of Asset Management*, *6*(6), 456–469. https://doi.org/10.1057/palgrave.jam.2240195

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M., & Viratyosin, T. (2020). The Unprecedented Stock Market Reaction to COVID-19. *The Review of Asset Pricing Studies*, *10*(4), 742–758. https://doi.org/10.1093/rapstu/raaa008

Banerji, G. (2020, September 15). *S&P 500 fastest recoveries following a bear market, from record high to new record* [Graph]. Wall Street Journal. https://www.wsj.com/articles/why-did-stock-markets-rebound-from-covid-in-record-time-here-are-five-reasons-11600182704

BlackBerry Industry Solutions. (2022). BlackBerry. Retrieved May 2, 2022, from https://www.blackberry.com/us/en/industries

Chaudhry, S., & Kulkarni, C. (2021, June). Design Patterns of Investing Apps and Their Effects on Investing Behaviors. *Designing Interactive Systems Conference 2021*. https://doi.org/10.1145/3461778.3462008

Chen, S. (2020). Forecasting Daily Stock Market Return with Multiple Linear Regression. *Mathematics Senior Capstone Papers*, *19*. https://digitalcommons.latech.edu/mathematics-senior-capstone-papers/19

Chung, J. (2021, February 18). *Melvin Capital Says It Was Short GameStop Since 2014*. WSJ. Retrieved May 5, 2022, from https://www.wsj.com/articles/melvin-capital-says-it-has-been-short-gamestop-since-2014-11613593854

Chunxi Liu, Li Su, Qingming Huang, & Shuqiang Jiang. (2011, July). News video story sentiment classification and ranking. *2011 IEEE International Conference on Multimedia and Expo*. https://doi.org/10.1109/icme.2011.6011900

Cieliebak, M. (2022). *Sentiment Analysis: Distinguish Positive and Negative Documents – SpinningBytes*. Spinningbytes. Retrieved March 5, 2022, from https://spinningbytes.com/sentiment-analysis-distinguish-positive-and-negative-documents/

Corporate Profile | AMC Theatres. (2022). AMC. Retrieved May 2, 2022, from https://investor.amctheatres.com/corporate-overview/default.aspx#

Droke, C. (2001). *Moving Averages Simplified*. Adfo Books.

Dunford, R., Su, Q., & Tamang, E. (2014). The Pareto Principle. The Plymouth Student Scientist, 7, 140-148.

Emerson, K. (2019, November 7). *An Introduction to Web Scraping for Research*. Research Data Services. Retrieved March 6, 2022, from https://researchdata.wisc.edu/news/an-introduction-to-web-scraping-for-research/

EV/EBITDA (Enterprise Multiple) by Sector/Industry 1995 – 2022. (2022, January 13). [Dataset]. https://siblisresearch.com/data/ev-ebitda-multiple/

Gandel, S. (2021, February 2). WallStreetBets says Reddit group hit by "large amount" of bot activity. *CBS News*. https://www.cbsnews.com/news/WallStreetBets-reddit-bot-activity/

Ghaeli, M. R. (2017). Price-to-earnings ratio: A state-of-art review. *Accounting*, *3*(2), 131–136. https://doi.org/10.5267/j.ac.2016.7.002

Giglio, S., Maggiori, M., Stroebel, J., & Utkus, S. (2021). The joint dynamics of investor beliefs and trading during the COVID-19 crash. *Proceedings of the National Academy of Sciences*, *118*(4). https://doi.org/10.1073/pnas.2010316118

Gill, K. (2021, January 29). *GME YOLO month-end update — Jan 2021* [Screenshot]. reddit.com/r/WallStreetBets. https://www.reddit.com/r/WallStreetBets/comments/l846a1/gme_yolo_monthend_update_jan_2021/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* (2016th ed.). MIT Press. https://www.deeplearningbook.org/

Gulcehre, C., Cho, K., Pascanu, R., & Bengio, Y. (2014). Learned-Norm Pooling for Deep Feedforward and Recurrent Neural Networks. *Machine Learning and Knowledge Discovery in Databases*, 530–546. https://doi.org/10.1007/978-3-662-44848-9_34

Gupta, V. (2018). Predicting Accuracy of Valuation Multiples Using Value Drivers: Evidence from Indian Listed Firms. *Theoretical Economics Letters*, *08*(05), 755–772. https://doi.org/10.4236/tel.2018.85052

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *06*(02), 107–116. https://doi.org/10.1142/s0218488598000094

Homepage. (2022). Reddit. Retrieved February 28, 2022, from https://www.redditinc.com/

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Institute for Memetic Research & Development. (2021). *brrr.money*. brrr.money. https://brrr.money/

Jennergren, L. P. (2008). Continuing value in firm valuation by the discounted cash flow model. *European Journal of Operational Research*, *185*(3), 1548–1563. https://doi.org/10.1016/j.ejor.2006.08.012

Jitmaneeroj, B. (2017). Does investor sentiment affect price-earnings ratios? *Studies in Economics and Finance*, *34*(2), 183–193. https://doi.org/10.1108/sef-09-2015-0229

Jung-Tae Jo, & SangHyunChoi. (2015, September). Sentiment Analysis of movie review for predicting movie rating. *Management & Information Systems Review*, 161–177. https://doi.org/10.29214/damis.2015.34.3.009

Jupyter Notebook. (2014). [Non-profit, open-source project for Python]. https://jupyter.org/

Ko, Y. (2012). A study of term weighting schemes using class information for text classification. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*. https://doi.org/10.1145/2348283.2348453

Kruschwitz, L., & Loeffler, A. (2005). *Discounted Cash Flow* (New title ed.). Wiley.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Loria, S. (2020). *TextBlob* (v0.16.0) [Processing textual data]. https://textblob.readthedocs.io/en/dev/index.html

Martínez Torres, J., Iglesias Comesaña, C., & García-Nieto, P. J. (2019). Review: machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, *10*(10), 2823–2836. https://doi.org/10.1007/s13042-018-00906-1

Mateus, B. C., Mendes, M., Farinha, J. T., Assis, R., & Cardoso, A. M. (2021). Comparing LSTM and GRU Models to Predict the Condition of a Pulp Paper Press. *Energies*, *14*(21), 6958. https://doi.org/10.3390/en14216958

McCabe, C., Banerji, G., & Frankl-Duval, M. (2021, January 11). *TikTok and Discord Are the New Wall Street Trading Desks*. WSJ. Retrieved May 5, 2022, from https://www.wsj.com/articles/tiktok-and-discord-are-the-new-wall-street-trading-desks-11610361004?mod=Searchresults_pos19&page=2

Mitchell, R. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web* (2nd ed.). O'Reilly Media.

Mohammad, S. (2016). A Practical Guide to Sentiment Annotation: Challenges and Solutions. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. https://doi.org/10.18653/v1/w16-0429

Nabi, J. (2021, December 10). *Recurrent Neural Networks (RNNs) - Towards Data Science*. Medium. Retrieved April 6, 2022, from https://towardsdatascience.com/recurrent-neural-networks-rnns-3f06d7653a85

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Natural Language Toolkit (NLTK) (3.7). (2001). [Open source Python library for NLP]. https://www.nltk.org/index.html

Ng, A. (2018). *CS229 Lecture notes* [Slides]. See.Stanford.Edu. https://see.stanford.edu/materials/aimlcs229/cs229-notes3.pdf

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Amsterdam University Press. http://neuralnetworksanddeeplearning.com/

Pagano, M. S., Sedunov, J., & Velthuis, R. (2021). How did retail investors respond to the COVID-19 pandemic? The effect of Robinhood brokerage customers on market quality. *Finance Research Letters*, *43*, 101946. https://doi.org/10.1016/j.frl.2021.101946

Pang, B., & Lee, L. (2008a). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pang, B., & Lee, L. (2008b). *Opinion Mining and Sentiment Analysis* (Vol. 2, Issues 1–2). Now Publishers. https://doi.org/10.1561/1500000011

Prentice, C. P. S. (2021, January 29). *Famed GameStop bull "Roaring Kitty" is a Massachusetts financial advisor*. Reuters. Retrieved December 8, 2021, from https://www.reuters.com/article/us-retail-trading-roaringkitty/famed-gamestop-bull-roaring-kitty-is-a-massachusetts-financial-advisor-idUSKBN29Y0AF

Provost, F., & Fawcett, T. (2013). *Data Science for Business*. Van Duuren Media.

Reddit. (2019, August 29). *overview for user*. Retrieved April 7, 2022, from https://www.reddit.com/user/DeepFuckingValue/

Saha, A., Malkiel, B. G., & Rinaudo, A. (2018). Has the VIX index been manipulated? *Journal of Asset Management*, *20*(1), 1–14. https://doi.org/10.1057/s41260-018-00102-4

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Scikit-learn. (2007). [Cumputer science for Python]. https://scikit-learn.org/stable/index.html

Sedkaoui, S. (2018). Machine Learning: a Method of Data Analysis that Automates Analytical Model Building. *Data Analytics and Big Data*, 101–122. https://doi.org/10.1002/9781119528043.ch6

Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine Learning Approaches in Stock Price Prediction: A Systematic Review. *Journal of Physics: Conference Series*, *2161*(1), 012065. https://doi.org/10.1088/1742-6596/2161/1/012065

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacology*, *10*(1). https://doi.org/10.1186/1471-2210-10-6

Su, Y., Cui, C., & Qu, H. (2022). Self-Attentive Moving Average for Time Series Prediction. *Applied Sciences*, *12*(7), 3602. https://doi.org/10.3390/app12073602

The Complete Toolbox For Investors. (2022). Finbox.Com. Retrieved March 26, 2022, from https://finbox.com/watchlist

The Economist. (2021a, July 6). *Are "meme stocks" harmless fun, or a threat to the financial old guard?* Retrieved February 5, 2022, from https://www.economist.com/the-economist-explains/2021/07/06/are-meme-stocks-harmless-fun-or-a-threat-to-the-financial-old-guard

The Economist. (2021b, July 6). *Are "meme stocks" harmless fun, or a threat to the financial old guard?* Retrieved February 5, 2022, from https://www.economist.com/the-economist-explains/2021/07/06/are-meme-stocks-harmless-fun-or-a-threat-to-the-financial-old-guard

UNITED STATES DEPARTMENT OF LABOR. (2020–2021). *Unemployment Insurance Weekly Claims Data* (Version 2020) [Dataset]. UNITED STATES DEPARTMENT OF LABOR. https://oui.doleta.gov/unemploy/claims.asp

Unwin, A. (2020). Why is Data Visualization Important? What is Important in Data Visualization? *2.1*. https://doi.org/10.1162/99608f92.8ae4d525

van Rossum, G. (1991). *Python* (3.10.4) [Programming language]. Python Software Foundation. https://www.Python.org/

Vasileiou, E., Bartzou, E., & Tzanakis, P. (2021). Explaining Gamestop Short Squeeze using Intraday Data and Google Searches. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3805630

WallStreetBets. (2012, January 31). Reddit.Com. Retrieved December 8, 2021, from https://www.reddit.com/r/WallStreetBets/wiki/faq

Appendix

The attachments are added separately to this research paper in a zip file due to their size. They include all the python notebooks used for data gathering, data preparation and modelling, as well as the gathered data from WallStreetBets, the sentiment analysis model output and the input data for each of the GRU and MLR models.