Capstone Project Planning Document

TEXT ANALYTICS OF THE GAMESTOP (GME) FRENZY ON REDDIT

By

Ng Wei Xiang

(18033167)

Bachelor (HONS) Information Systems (Honours) Data Analytics

Supervisor: Dr Mohammad Dabbagh

Date: 23 November 2021

Table of Contents

1. Introduction 3
   1.1 Background 3
   1.2 Problem Statement 6
   1.3 Project Aim and Objectives 8
   1.4 Project Scope 8
2. Literature Review 10
   2.1 Stock market and the concept of short squeeze 10
   2.2 GameStop 14
   2.3 Analytical works on the stock market 16
   2.4 Analytical works on GameStop 18
      2.4.1 Sentiment Analytics on GameStop 18
      2.4.2 Predictive Analytics on GameStop 18
      2.4.3 Other Analytics on GameStop 20
   2.5 Conclusion and Remarks 21
3. Research methodology 25
   3.1 Research Objectives 25
   3.2 Project Planning (CRISP-DM) 29
      3.1.1 Phase 1: Business Understanding 30
      3.1.2 Phase 2: Data Understanding 34
      3.1.3 Phase 3: Data Preparation 44
      3.1.4 Phase 4: Modelling 48
      3.1.5 Phase 5: Evaluation 53
      3.1.6 Phase 6: Deployment 55
   3.3 Summary 58
4. Work Plan 61
   4.1 Work Activities 62
   4.2 Flow Chart 68
   4.3 Gantt Chart 69
5. References 70
   5.1 List of Tables 70
   5.2 List of Figures 71
   5.3 Bibliography 72

1. Introduction

This section will be discussing about the general knowledge and background of the short squeeze event. Due to this being a real-world event that happened months ago, some knowledge and background story will be presented to provide a better understanding of the situation. Then, the problem statement will be identified based on the background. This leads to the project aim and objective to address the problem statement. Finally, a project scope will be defined to show and visualize how and what would be done.

1.1 Background

In January 2021, there happens a short squeeze on the stock of GameStop (GME) that causes a major financial consequence for some hedge funds and short sellers which the story all starts and is about GameStop. GameStop is an American chain video game store that operates in a B2C and brick-and-mortar manner. In the recent years with the digitalization of businesses, GameStop has struggled to keep up with the competition from digital distribution services platforms such as Epic Games Store, Steam etc by offering a competitive solution on bringing their business online. With the COVID-19 pandemic hitting the United States in 2019, the situation of GameStop went worse when the people visiting their stores in person saw a decreasing trend. As a result, the stock price of GameStop declined which leads to many institutional investors and hedge funds selling short on the stock. According to Forbes (Ponciano, 2021), in January 2021, an approximate of 140% of GameStop's public float had been sold short, which means that some shorted shares, had been re-lent and shorted in the market again.

Then, the 2nd entity of this short squeeze arrives – the subreddit community or also known as "r/wallstreetbets". This community is known for a place for discussion around high-risk-and-return stock transactions. The community in r/wallstreetbets believed that the company was highly undervalued, and with such huge number of shares being shorted, a short squeeze could potentially be triggered where short sellers had to capitulate and cover their positions at large losses. Before the short squeeze even happened, there is a person named Keith Gill or goes by the Reddit username of "DeepFuckingValue", showed high interest of the stock in which he purchased 53000 USD worth of call options in GameStop's stock in 2019 and saw a flip of 900 times when his account was worth at 48 million USD by 27th of January 2021 (Esguerra, 2021).

The short sellers which is the 3rd entity in this short squeeze are mostly consists of hedge funds and activist short seller. The entities include Melvin Capital, Citadel LLC, Point72 Asset Management, Citron Research, and D1 Capital Partners. Short sellers are those who opens a position by borrowing shares of a stock or other assets when they believe they will decrease in value. The short sellers will then sell away these borrowed shares to buyers that are willing to pay for the market price at that moment. Essentially, short sellers are betting that the price will continue decrease and so they can purchase them at a lower cost and repay the borrowed shares. According to The Wall Street Journal (Chung, 2021), Melvin Capital had lost 53% of its investment shorting the stock by the end of January. At the same time, one of the activist short sellers, head of Citron Research, Andrew Left, claimed to close the position at a total loss. In a further interview, Andrew shared that the company covered most of its short positions in the range of $90 per share at a loss of 100% while retaining a small manageable position (Citron Research, 2021).

Lastly, the last important entity in this short squeeze is Ryan Cohen, which is the former CEO of the online pet food retailer – Chewy. In September 2020, Ryan Cohen showed a significant investment in GameStop and joined the company's board of directors. This news quickly gained attention, leading to some who believe that the stock was undervalued and will rise with the presence of Ryan in helping GameStop improve (Stewart, 2021).

With the entities laid out clearly, the story of this short squeeze can now be explained in a chronological order. In 2019, a person named Keith Gill or known as "DeepFuckingValue" in reddit started buying long on GameStop because he believed that GameStop is highly undervalued. Soon enough, his portfolio went from $50,000 to $48 million and slowly caught the attention of the subreddit community "r/wallstreetbets". In September 2020, the former CEO of Chewy.com, Ryan Cohen, showed a significant investment and announced that he is joining the board in GameStop. This gained more attention when the ex-CEO of one of the e-commerce giants decided to join a company that is struggling in modernizing their sales. Despite Ryan Cohen joining the board in GameStop, hedge funds like Melvin Capital and Citron Research still believes that GameStop will not be performing better anytime soon and decided to bet by shorting the stock of GameStop. With that, the community of the subreddit group decided to join in and start this short squeeze buy buying long and act otherwise against the hedge funds which the result was seen around January 2021, where the stock price of GameStop skyrocketed to an all-time high of $483. The outcome of this short squeeze was seen where the hedge funds had lost significantly, while the community of the subreddit group brings back huge gains and profit from this short squeeze.

On January 28, 2021, one of the largest stock trading platforms, Robinhood, halted purchases of GameStop and several other stocks like AMC Theaters and Nokia Corporation which was also the other short squeezed stocks (Winck, 2021). This means that the customers of Robinhood could no longer open new positions in the stock, although they could still close their positions. Soon enough, other brokerages like eToro followed the suit. After the markets closed that day, Robinhood announced that they would begin allowing "limited buys" starting the following day, although it was unclear what "limited buys" entail (Robinhood, 2021). Then as of January 26, 2021, Robinhood was still having limits on the trading of GameStop, AMC, and Blackberry stocks. Few days later at the 1st and 2nd of February 2021, the stock price of GameStop plummeted substantially, which saw a drop of more than 80 percent from its all-time high peak price. 60 percent of their share values was lost on February 2 (Financial Times, 2021), while a report showed a total of $27 billion was wiped following the plummet of stock price (Lipschultz, GameStop Rout Erases $27 Billion as Reddit Favorites Tumble, 2021). CNN reported that the drop of price was partly due to the restriction imposed by Robinhood and other brokers on the stock which the stock price seem to continue the declination during the week (Monica, 2021) to $63 per share.

With the active discussion and hype still going around the community in r/wallstreetbets, a resurgence of the price of GameStop was seen on 24th February 2021. The share price doubled in the final 90 minutes in the trading hour, which saw a final closing price at $91.70 that represents a 104 percent gain. The gains even continued in the after hours, to nearly 100 percent. Soon on March 8, 2021, the price continued rising for 41 percent to $194.50 with ¼ of the stocks being shorted at this time (Lipschultz, 'Reddit Raider' Favorite GameStop Soars on Latest Cohen Push, 2021). On the next day 9th March 2021, the stock surged to its highest point since January and closed at $246.90. on March 10, 2021, the price rose again to $348.50 before getting paused for volatility and finally dropping 40 percent. 2 weeks later at March 24, 2021, the stock price plummeted 34 percent to $120.34 after the earnings report of the company was released and the company's plan for issuing a new secondary stock offering.

1.2 Problem Statement

Till date, the trend of short squeeze seemed to flatten with the price of the stock getting stable after the event. However, the price did not go back to the price before the short squeeze which is between $10 to $20. At current date (15th September 2021), the price of GameStop's stock sits at $199, which is 10 times higher as compared. At the same time, the conversation and trend in the subreddit did not seem ending anytime soon with the community highly encouraging to buy and hold their positions.

Before this short squeeze event on GameStop, there are also several other short squeezes happened in the history that was worth an attention. In 2008, there was a moment that Volkswagen being the most valuable company in the world with a share price of more than €1000, was revealed that Porsche had gain control of 74 percent of Volkswagen's voting shares by buying most of the circulating stocks that Volkswagen had. The short squeeze took place the event took place during the global financial crisis and when an approximate of 12.5 percent of Volkswagen's stock was on loan to short sellers because they were viewed as a high probability bankruptcy candidate. When the market opened the next day after Porsche's announcement, those short sellers rushed to minimize their positions by buying in more stocks, which caused the share price to inflate more. Then on 27th October 2008, Volkswagen's share price opened at €348 and closed at €517. That time, the stock peaked at €999 per share with short-selling costs were estimated at $30 billion (Mox reports, 2018).

By comparison, the short squeeze on GameStop and Volkswagen had a few similarities and difference. The first is that they both were listed as one of the bankruptcy candidates. Secondly, they both were shorted for a common reason, which is a significant reduction of sales. However, looking at the differences, the most obvious difference is the absence and present of social media between both events. Volkswagen in 2008 does not receive wide attention by the public due to the absence of social media while GameStop in 2021 had gained much attention due to the active participation and discussion on Reddit and exposure on other social medias. Next, we can also see that Porsche's short squeeze took place during the financial crisis period, while GameStop's short squeeze occurred during the global Covid-19 pandemic. Most importantly, there is no stops or halts from any parties like brokerages during the Volkswagen short squeeze but there were a lot during the GameStop short squeeze.

Table 1: Differences and similarities between short squeezes

	Volkswagen	GameStop
Differences	- Absence of social media - During global financial crisis - No halts by any other parties	- Presence of social media - During global Covid-19 pandemic - Saw halts by brokerages and platforms
Similarity	1. Bankruptcy candidates 2. Significant reduction in sales 3. Occurred during global crisis	

These 2 short squeeze examples raise a question that what exactly is affecting a company's stock price and who are manipulating the prices all the time. All this time for people without a strong financial knowledge or background, it is hard for people like us to understand and explain how money, funds, or the stock market operates. Some says that it is controlled by the hedge funds, or the brokers in Wall Street, while some say that it is controlled by the politicians. However, the GameStop short squeeze proves otherwise. A group of people actively discussing about a company on a social media platform, then put in effort and purchased stocks, end up beating the people in or behind Wall Street and claim the success. So, who is really controlling the variation of price and stock, and how the problem occurs are what I would like to answer.

1.3 Project Aim and Objectives

The aim of this project is to investigate the social media side of the short squeeze on GameStop. This may help us understand of digitalization affects our daily life and the market and the difference between them. With social media being the common communication medium of people around the globe, it would be interesting to see how people think based on their comments on them. With that, the main objective is to understand the psychological behavior of the community on the short squeeze to know better on their thoughts. To achieve the aim, the following objectives are proposed:

- To investigate the sentimental values of the community about the short squeeze alongside with the keywords by generating a word cloud
- To build a clustering model and cluster the types of users on the reddit subcommunity based on their comments.
- To build a predictive model and predict the change of stock price of GameStop in the future

1.4 Project Scope

With the objectives laid out proposed, this project will be split into multiple stages to assure that the expected outcome and result could be achieved.

At the initial stage, there will be more understanding of the business case and the exact types of data will need to be identified. The project research methodology will be following the CRISP-DM methodology which the next step will be data mining based on the understanding of the case and data. Once text data are mined, the next step according to the methodology is understanding the data. Data understanding includes identifying what is the meaning of this data, where to use it, and how to use it. Once we had understood our data thoroughly, the next step will be preparing the data. Data preparation includes data cleaning, manipulation, creation, and more based on our objectives stated clearly above. Once the data was ready, several models will be generated for each objective to get a best performing model. Once done, it moves on to the evaluation stage where the best model is selected and evaluated. Understanding how the model brings the best accuracy is key before we proceed for deployment.

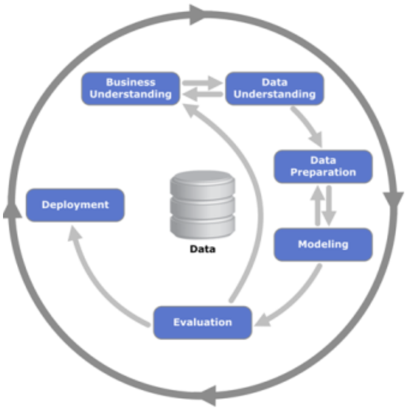


Figure 1: Image representation of the CRISP-DM methodology (Rodrigues, 2020)

The project scope is designed in a way that each stage will be covered carefully and in detail to achieve optimal result and performance in the end of the project. To obtain relevant data, python scripts will be written to scrape an abundance of text data on the comments of posts in the subreddit page “r/wallstreetbets”. Then, the mined data will be stored and understood thoroughly, then cleaned or re-scrape to obtain ideal data for each objective. With that, the data will be prepared with necessary steps such as cleaning, manipulation, creation, binning or more to ensure that our work is relevant to the research. Modelling in this research includes supervised modelling like regression, NLP programming, word cloud generation and more. These models are necessary to help achieve the objectives and provide better visualization of our results. Once the result of the models is optimal, the results will be deployed and explained further to display the insights.

2. Literature Review

The purpose of this literature review is to better understand about the stock market, social media, and the people's behavior on them. This literature review will provide a thorough understanding about them for deeper and accurate analytics in the future. The approach and strategy taken to perform this literature review is from 3 directions: stock market related research, GameStop frenzy related research and lastly predictive or sentiment analytics related research. These 3 directions were aimed specifically because it is what makes up the GameStop frenzy and belongs under the scope of this research topic. Thus, extensive literature review is required to be performed from these 3 points to better understand past works and potential solutions.

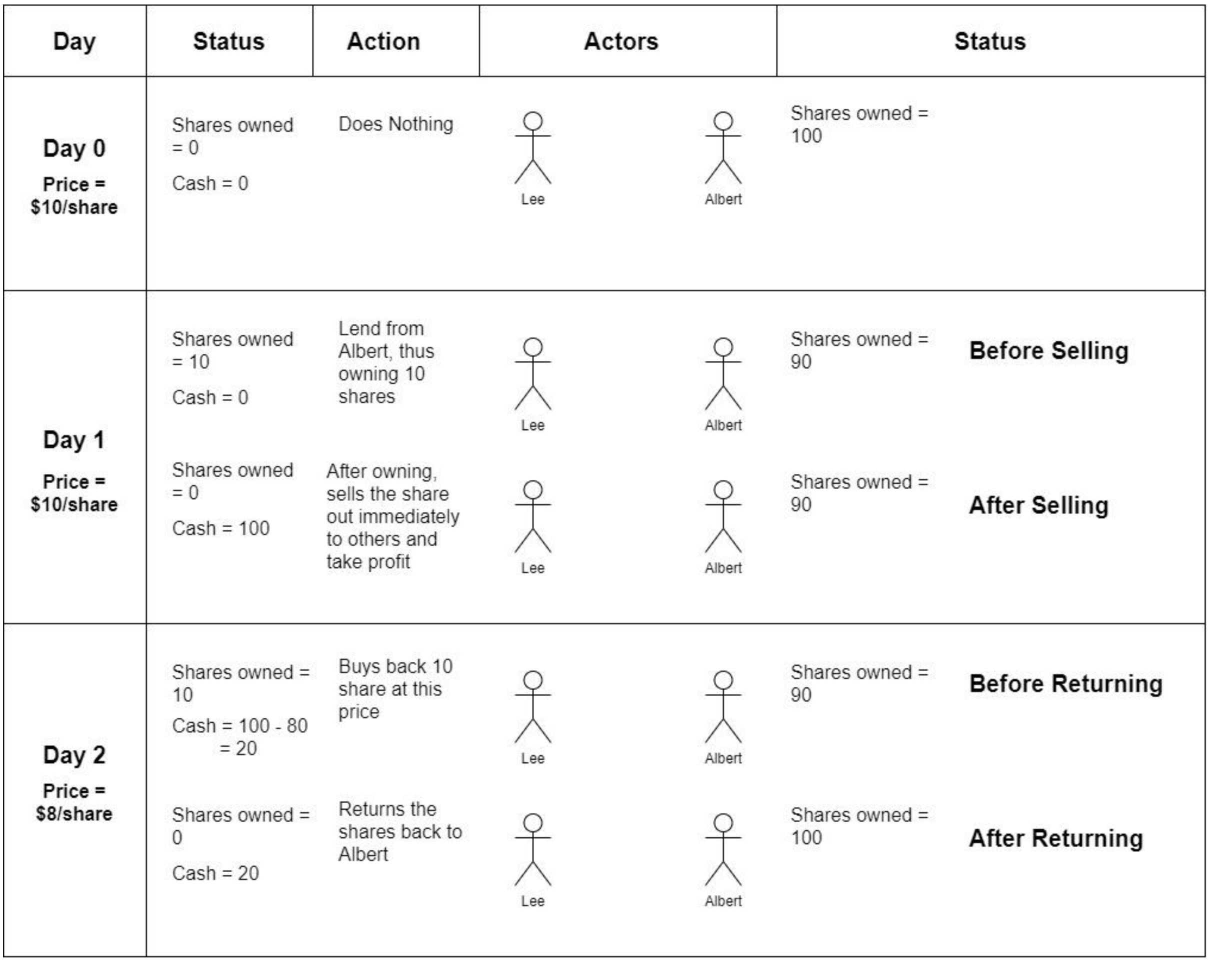
2.1 Stock market and the concept of short squeeze

To start the literature review, it is always better to have an understanding and background concept on stock market and short squeeze. According to Wikipedia (Wikipedia, n.d.), the stock market was defined as the aggregation of buyers and sellers of stocks(shares), which represent ownership claims on businesses. Any investment or trading activities to be done on the share market are mostly done via stockbrokerages and electronic trading platforms. The purpose of stock market existing is to act as one of the methods to help companies to raise additional money for expansion. In exchange, these companies will sell shares of the ownership of the company in the public market. Making the concept simple, an example will be demonstrated to simply the understanding. For example, Mr Andrew owns a lemonade stall and was generating some profits well. However, he noticed an opportunity of selling his lemonade down this street would potentially bring in more profit, but he does not have sufficient money(capital) to open another stall. Thus, the potential solution, is to break up his business into multiple shares, and sell it to other people that has the money and would like to own a margin of shares in his business. This is the purpose of stock market and companies listing them on a nutshell.

With that, once Mr. Andrew's lemonade business is listed and traded, he receives the money(capital) to expand his business while the investors owned some share of his business. Few months later, Mr. Andrew's decision was proven to be right, which his business was seen growing, and the investors are willing to invest in more money to have more shares on his business. The demand is growing. However, a year later, a thunderstorm destroyed one of his stalls, and his profits were seen dropping. The investors were worried and rushed to sell their shares to other investors because they do not want to bear the losses. Now, the supply demand is dropping, while supplies are growing. These buys and sells are the fundamental reasons of the fluctuation of share prices in the stock market, which was all about supplies and demands.

According to Investopedia (Hayes, 2021), the price of shares on a stock market can be set in multiple ways, with most common one being an auction process where buyers and sellers place their bids to buy or sell the shares. “Bid” is where the person wishes to buy the share at the price, and “offer” is where the person wishes to sell the share at the price. And when the “bid” and “offer” coincides with each other, a trade or transaction is made. The overall stock market is made up by millions of traders or investors, that has different opinion of what the price of the stock should be at that moment. The transactions between the traders bidding and offering the stock, is what makes the price of a stock to fluctuate every minute, hour, day, week, month, and year. The price of a stock increases when the demand(bid) surpasses the supply(offer) and decreases when the supply(offer) surpasses the demand(bid). In the financial market, investors also use the term “long” and “short” to represent their opinion on the stock on whether they will rise or fall. “long” means that they believe that the stock will have more demand than supply, thus believing that the price will go up. Inversely, “short” means that the investors believe that the stock will have more supply than demand, thus driving the price down.

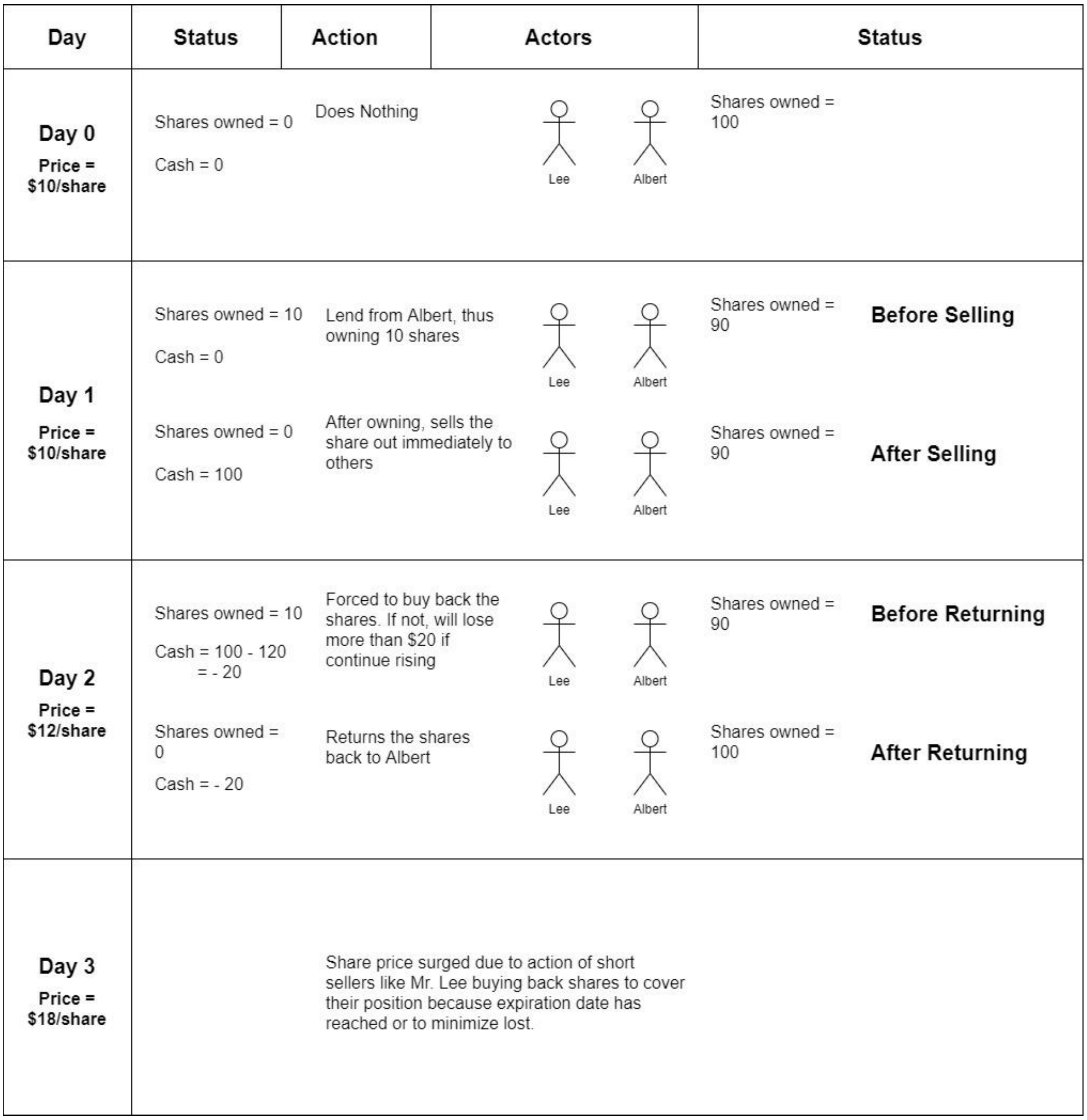
Before that, we will need to understand the concept of another term called “short selling”. Short selling is a practice where investors is looking down on the price of a stock and trying to profit while the price is declining. They normally borrow a number of shares of a company, wait for the price to decline, and buy them back later at a lower price. They profit by spending less in buying back the shares they earned during the process of selling them. Using back the previous example, Mr. Lee would like to sell short on Mr. Andrew’s lemonade business. To do that, he lent 10 shares at a current price of $10 from Mr. Albert who owns some shares of the lemonade business. The thunderstorm hit 1 day later and damaged the lemonade business, which is a bad news to all investors and triggers a wave of share selling. The share of the business fell to $8 per share the day after the thunderstorm and Mr. Lee buys back 10 shares immediately. Thus, Mr. Lee gains a profit of $20 this way: lent 10 shares at $10 which is worth of $100 at day 1. A bad news arrive at day 2, the stock price drops, Mr. Lee buys back 10 shares at $8 which is worth of $80 and return to Mr. Albert. With that, Mr. Lee owned the same number of shares which is 10 of them, by buying them later at a lower price while taking home a profit of $20 which could be visualized in figure 2.



Day	Status	Action	Actors	Status
Day 0 **Price = $10/share**	Shares owned = 0 Cash = 0	Does Nothing	Lee Albert	Shares owned = 100
Day 1 **Price = $10/share**	Shares owned = 10 Cash = 0	Lend from Albert, thus owning 10 shares	Lee Albert	Shares owned = 90 **Before Selling**
	Shares owned = 0 Cash = 100	After owning, sells the share out immediately to others and take profit	Lee Albert	Shares owned = 90 **After Selling**
Day 2 **Price = $8/share**	Shares owned = 10 Cash = 100 - 80 = 20	Buys back 10 share at this price	Lee Albert	Shares owned = 90 **Before Returning**
	Shares owned = 0 Cash = 20	Returns the shares back to Albert	Lee Albert	Shares owned = 100 **After Returning**

Figure 2: simple demonstration of the short selling concept

According to Investopedia (Mitchell, 2021), short squeeze was defined as an unusual condition that triggers rapid rising prices in a stock. All short sales will have an expiration date, and when the stock rises unexpectedly, the short sellers will be forced to buy at a higher price and pay for the difference. When short sellers exit their positions by buying more share, the coincidence of these with other short sellers will drive the price higher because all of them had to buy to cover their losses. With short sellers panicking and buying in more shares, combined with the new buyers, is what causes a surge in the share price and therefore short squeeze occurs which could be visualized in figure 3.



Day	Status	Action	Actors	Status	
Day 0 **Price = $10/share**	Shares owned = 0 Cash = 0	Does Nothing	Lee Albert	Shares owned = 100	
Day 1 **Price = $10/share**	Shares owned = 10 Cash = 0	Lend from Albert, thus owning 10 shares	Lee Albert	Shares owned = 90	**Before Selling**
	Shares owned = 0 Cash = 100	After owning, sells the share out immediately to others	Lee Albert	Shares owned = 90	**After Selling**
Day 2 **Price = $12/share**	Shares owned = 10 Cash = 100 - 120 = - 20	Forced to buy back the shares. If not, will lose more than $20 if continue rising	Lee Albert	Shares owned = 90	**Before Returning**
	Shares owned = 0 Cash = - 20	Returns the shares back to Albert	Lee Albert	Shares owned = 100	**After Returning**
Day 3 **Price = $18/share**		Share price surged due to action of short sellers like Mr. Lee buying back shares to cover their position because expiration date has reached or to minimize lost.			

Figure 3: simple visualization of the concept of short squeeze

Now with an understanding on stock market on short squeeze concept, we should have a clearer picture on what is happening and how. We shall continue the literature review to next section to understand the research done about GameStop.

2.2 GameStop

GameStop (ticker symbol: GME) is an American listed company on the New York Stock Exchange (NYSE) which serves as a video game, consumer electronics, and gaming merchandise retailer. GameStop has 4816 stores in the world with a vast majority of them operating in the United States, while the remaining are seen operating in Canada, Australia, New Zealand, and Europe under different brands. It is founded at 1984 and had been in operation for around 37 years in this industry which was seen as a go-to place for the gamers. The arrival of the Covid-19 pandemic at the end of 2019 and early 2020 impacted everyone's lifestyle which a significant amount of population started the transition to be working from home. Kerckhoven and O'Dubhghaill (2021) mentioned that this in turn facilitated people to look for alternatives to spend their leisure time, on latest trends to go after and ventures to participate (Kerckhoven & O'Dubhghaill, 2021).They further explianed by mentioning about the implmentation of the lockdown and various quarantine measures helped increase the interest and attention of investments and the stock market. Furthermore, due to the market volatility which a depression of share prices occurred at March of 2020, quickly attracted attention of more novice traders to gain more profits by actively trading on online commission free brokerages such as RobinHood where over 1 million of new online brokerage accounts were created at the first quarter of 2020 alone (Burnette, 2021). It was also cited by Burnette (2021) that total market trades in July and August 2020, 25% of them were accounted by retail investors (individual traders).

Fast forward to early January 2021, no one ever though that these novice traders would quickly be a force to reckon in the GameStop frenzy. One of the traders that is active on "r/WallStreetsBets" noticed that hedge funds such as Melvin Capital and Citron Research had been shorting the stock price of GameStop. He then launched the idea of buying the stocks long as an action against them so that the share price could rise back which slowly gained attention more attention by the community in the subreddit page where the novice traders joined in the movement and bought shares of GameStop. Kerckhoven and O'Dubhghaill (2021) also mentioned that the individuals had been buying the stock with the intention of holding them until the hedge funds had to exit their short positions. This movement gained more traction and attention which the price of GameStop reached an all-time high of $483 on 28th January 2021.

According to Burnette (2021), the 3 main factors that led the GameStop short squeeze could be due to the change of demographics of retail traders, the presence of technology and internet within retail investors, and also social media. It is said that male investors have different risk tolerance when compared to women. A study by (Hibbert, Lawrence, & Prakash, 2008) found that men are more likely to invest in volatile assets like stocks when compared to women. At the same time, according to economists and financial analysts, risk-taking increases as age gets younger where the younger investors tend to hold on more risky stocks. A study by (JianWei, 2015) showed that 88% of the online traders spend their time on the internet daily while 66.2% of non-online traders spend their time online daily. At the same time, it was seen that out of the online traders, 62.5% of them are men as compared to women at only 37.5% which explains the findings of Burnette (2021) that the demographic change of retail traders to be mostly dominated by men. With that, we can say that online traders can have easier access to information and sharing of opinions on the internet as they are more active and comfortable communicating online. The presence of technology was also seen especially important where the emergence of 0% commission fee brokerage apps like RobinHood, Webull, and eToro became the go-to platforms for stock trading for retail investors. Lastly, social media was also mentioned as one of the factors according to Burnette (2021). Research by (Lucey & Dowling, 2005) showed that mood and sentiments could be an efficient tool in decision making and showed its significant role in stock trading. This result acted as support to show how social media is influencing the retail investors' sentiment during their decision-making process on stock market. With Reddit being the powerhouse and the main camp of the retail investors to exchange ideas during the GameStop frenzy, more or less the social media site had showed the power of a social media in spreading and affecting sentiments of their users. Another evidence of the power of social media was seen where a higher social media coverage, can predict a higher increase in volatility return and trading activity (Peiran, Andre, & Ansgar, 2020).

It is seen that the GameStop frenzy is not something that was build up from nothing or anything minor. The conclusion of the study done by (Umar, Yousaf, & Zaremba, 2021) showed evidence that the retail investors on Reddit did influence the price of the stock of GameStop and had stimulated the shift of momentum towards other stocks too. Sufficient research studies and evidence had showed common agreement that the cause of this frenzy was due to the presence of a series of worldwide event. Starting from the Covid-19 pandemic which encouraged people staying at home made people spending more time on the Internet to pass their leisure time while looking for trends to pass their boredom. At the same time, the presence of online no-commission brokerage platforms further encouraged the participation of young active male netizens to open their accounts to start trading. Lastly, with the power of social media, people exchange information actively on social media sites, or in this case Reddit, one could easily affect other's opinion and sentiment which we saw a result of the active discussion in "r/WallStreetsBets" and soon the short squeeze of GameStop.

2.3 Analytical works on the stock market

Before we start with any analytics work, some researches do provide precautions and suggestion to obtain better results in analytics. For example, a paper by (Carvajal, 2021) is the perfect evidence in showing the importance of sufficient data where his paper ended in model that is insufficient to predict stock prices even with different approaches. He further explained that the possible reason behind the failure of prediction could be due to the features used was not strong enough. (Awate & Nandwalkar, 2019) did a survey to investigate and compare various methods for stock prediction. In the end, a conclusion and lesson were learned was to obtain proper dataset as training input while apply careful data extraction and normalization if needed to help achieve better accuracy. This paper and the paper by Carvajal (2021) equally showed importance of obtaining good data with sufficient data cleaning and manipulation as well as feature extraction so that an optimal accuracy can be obtained during predictive model building.

With some precautions and advice in mind, some papers also did suggest that a preparation of an enhanced dictionary of positive and negative words also help in obtaining better analytical results. In the paper done by (Joshi, Bharati, & Rao, 2016), they build their own positive and negative word dictionary with an addition of financial specific words and its sentiment with McDonald’s research (McDonald & Loughram, 2015). However, (Danqi, Charles, Valerie, & Xiaoyan, 2021) took a different approach when it comes to dictionary preparation than the usual reference and usage of McDonald’s research. They decided to build their own dictionary to better capture the lingo expressed on “r/WallStreetBets” by stripping away punctuations and numbers, then removal of stop words and finally lemmatizing and tokenizing words to obtain list of unique words. They then classify this list of words as positive, neutral, and negative manually. These 2 papers showed a really good work and example as to how to prepare dictionaries of positive and negative words for a better performing model. Although both papers did not show a comparison on the effect of the difference in building an enhanced dictionary, however they should contribute to a better model when past works had proved that a positive result was obtained by both papers. Another paper by (Belo, Erker, & Koehler, 2021) ended with a conclusion that says the models need to be trained with a financial vocabulary for better accuracy. These papers provided perfect example and reminder on the importance of preparation an enhanced dictionary with financial vocabulary.

Upon research through the Internet, many research and studies could be obtained to understand further about analytical information on the stock market or GameStop itself. One of the most cited work showed that opinions transmitted by investment social media like “Seeking Alpha” really help their users to make more informed decision where sentiment provides a strong and positive predictive ability where the future stock returns and earnings are predicted strongly (Hailiang, De, Hu, & Hwang, 2011). At the same time, research done by (Mehta, Pandya, & Koetcha, 2021) showed the sentiment of news can also successfully predict stock price where if it is a positive news, the price is likely to grow higher. They also showed that the Long-Short Term Memory (LSTM) machine learning method has higher accuracy in predicting the movement of stock price with sentiment polarity. (Nti, Adekoya, & Weyori, 2020) showed Multi-Layer Perceptron (MLP) is one of the commonly proven working method in predicting stock prices using sentiment. Within their research, they also suggested their framework and processing of data while obtaining a final accuracy of 77.12% with a combination of dataset by using the MLP algorithm. Another paper implemented text classification with different methodologies to predict the trends of the stock price (Joshi, Bharati, & Rao, 2016). They implemented TF-IDF scheme to represent the text documents and compared the performance between Random Forest, Naïve Bayes, and SVM classifier. They obtained a result that shows random forest and SVM classifier performs better than Naïve Bayes in a general view for classification of text in predicting stock trend with news sentiment.

In a general view, it was seen that different authors have their methods in performing their analytics to achieve their objectives. It seems like all the papers does not have a unified feature that can be referenced and followed, which is normal as people have different approach and objectives which results in different process. However, there are still some take home messages which was inspired by their works. For example, both (Awate & Nandwalkar, 2019) and (Carvajal, 2021) brought a lesson that shows the importance of having sufficient data and performing necessary steps for data cleaning, manipulation, and extraction. They showed the effects of having proper a dataset that was cleaned and manipulated accordingly will contribute to prediction models that are more accurate. This was further proven by the work done by (Hailiang, De, Hu, & Hwang, 2011) where they obtain their data from investment social media instead of the generic social media sites like Facebook, Twitter, or Reddit and resulted with a model that predicts positively.

2.4 Analytical works on GameStop

With sufficient understanding on background knowledges and analytical works on the stock market, the combination and investigation of the overlap of the 2 previous sections with GameStop leads to another question to be addressed, which is "What analytical work was performed to investigate the frenzy on GameStop?". Thus, it is broken down and organized according to the types of analytics performed. This section of literature review should be able to provide answer and insights to this question.

2.4.1 Sentiment Analytics on GameStop

The paper by (Belo, Erker, & Koehler, 2021) implemented BERT+ classifier for classifying text into their respective sentiment group showed a poor accuracy training dataset and resulted with insufficient evidence that sentiment analytics on Twitter is able to help financial analytics. However, they still observed that there is a high correlation of Twitter posts to the share price of GameStop. Another sentiment analytics work on GameStop done by (Long, Lucey, & Yarovaya, 2021) showed inspiring methodology and interesting results. They started with 4 hypotheses to investigate the extent and intensity of sentiments extracted from "r/WallStreetBets", which in general is about the tone and number of comments. Unlike other papers where the researchers employ their own dictionary, Long, Lucey, and Yarovaya used ready python packages such as "Text2emotions" and "SentimentIntensityAnalyser" to obtain sentiment and feelings while equipping the wavelet coherence framework to assess the relationship between sentiment and price change. In the end, a conclusion was seen where fear and sad are the most common tones which contrasts with the popular believe of anger. The research by (Danqi, Charles, Valerie, & Xiaoyan, 2021) also applied similar approach whereby they also capture the overall tones of the data for sentiment analytics.

The research by (Anand & Pathak, 2021) adopted valence shifters and ngram analysis for tone quantification during the preparation of sentiment dataset which was proven that to improve the accuracy over unigram analysis when included in the tone quantification process.

2.4.2 Predictive Analytics on GameStop

For predictive modelling, different approaches were seen from all papers which they had their own justification while some papers adopted multiple modelling techniques and obtain the best performing model through model comparison upon a common metrics such as precision and recall. However, the literature review about the predictive analytics could be segmented into 2 parts which is by their general approach. When they are trying to predict the price of GameStop's stock, is was observed that they prefer linear regression while if they are trying to predict the rise or fall of GameStop's stock, they prefer classification techniques such as Long Short-Term Memory (LSTM), Random Forest, Decision Tree, Naïve Bayes etc.

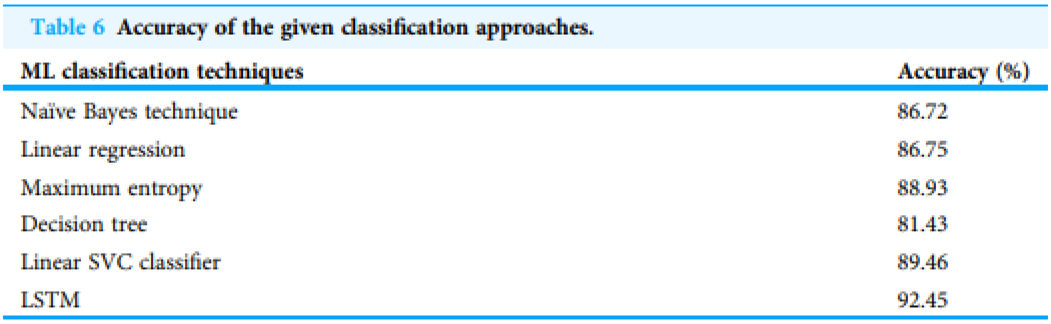
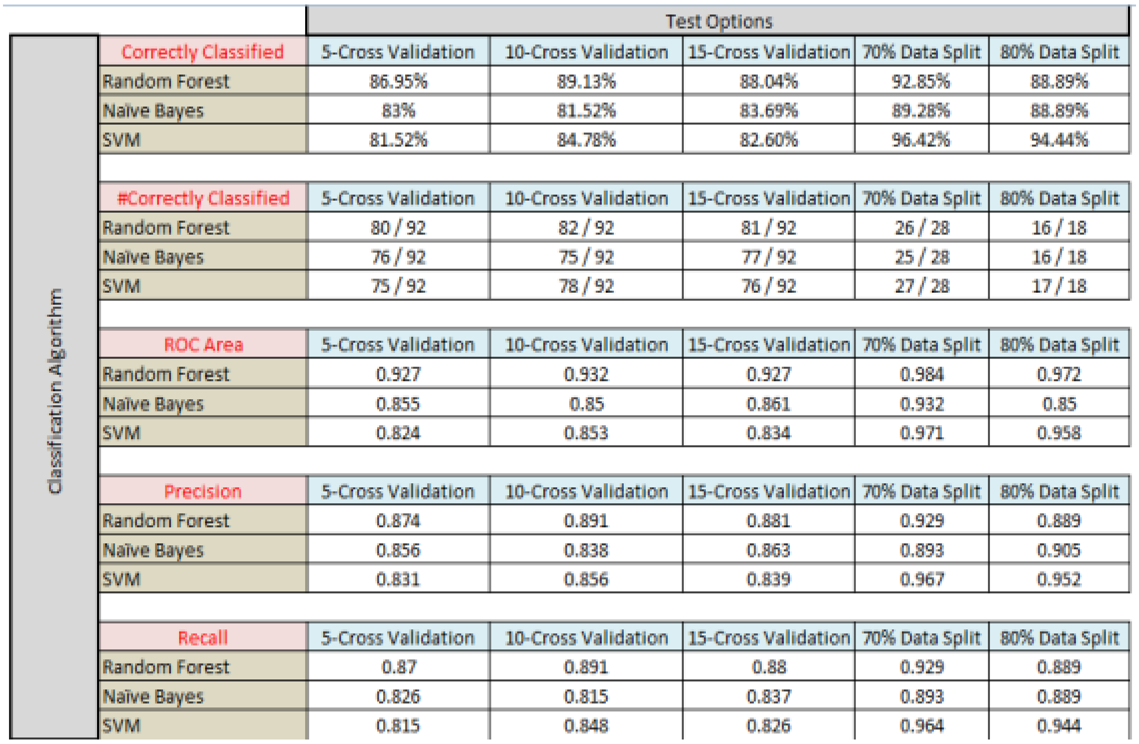


Table 6 **Accuracy of the given classification approaches.**

ML classification techniques	Accuracy (%)
Naïve Bayes technique	86.72
Linear regression	86.75
Maximum entropy	88.93
Decision tree	81.43
Linear SVC classifier	89.46
LSTM	92.45

Figure 4: Comparison of accuracy between different models for predictive modelling *(Mehta, Pandya, & Koetcha, 2021)*



		Test Options				
Classification Algorithm	Correctly Classified	5-Cross Validation	10-Cross Validation	15-Cross Validation	70% Data Split	80% Data Split
	Random Forest	86.95%	89.13%	88.04%	92.85%	88.89%
	Naïve Bayes	83%	81.52%	83.69%	89.28%	88.89%
	SVM	81.52%	84.78%	82.60%	96.42%	94.44%
	#Correctly Classified	5-Cross Validation	10-Cross Validation	15-Cross Validation	70% Data Split	80% Data Split
	Random Forest	80 / 92	82 / 92	81 / 92	26 / 28	16 / 18
	Naïve Bayes	76 / 92	75 / 92	77 / 92	25 / 28	16 / 18
	SVM	75 / 92	78 / 92	76 / 92	27 / 28	17 / 18
	ROC Area	5-Cross Validation	10-Cross Validation	15-Cross Validation	70% Data Split	80% Data Split
	Random Forest	0.927	0.932	0.927	0.984	0.972
	Naïve Bayes	0.855	0.85	0.861	0.932	0.85
	SVM	0.824	0.853	0.834	0.971	0.958
	Precision	5-Cross Validation	10-Cross Validation	15-Cross Validation	70% Data Split	80% Data Split
	Random Forest	0.874	0.891	0.881	0.929	0.889
	Naïve Bayes	0.856	0.838	0.863	0.893	0.905
	SVM	0.831	0.856	0.839	0.967	0.952
	Recall	5-Cross Validation	10-Cross Validation	15-Cross Validation	70% Data Split	80% Data Split
	Random Forest	0.87	0.891	0.88	0.929	0.889
	Naïve Bayes	0.826	0.815	0.837	0.893	0.889
	SVM	0.815	0.848	0.826	0.964	0.944

Figure 5: Comparison between different classification models for predictive models *(Joshi, Bharati, & Rao, 2016)*

Starting with predictive classification modelling, (Mehta, Pandya, & Koetcha, 2021) did a comparison between the performance of different predictive techniques and tabulated it. Figure 4 showed that Long Short-term Memory (LSTM) performs the best out of all of the 6 machine learning classification techniques for predictive modelling. A similar approach was seen where (Joshi, Bharati, & Rao, 2016) compared the performance between Random Forest, Naïve Bayes, and SVM where a result was seen where SVM and Random Forest performs better than Naïve Bayes in figure 5. Another example by (Nti, Adekoya, & Weyori, 2020) applied Multi-Layer

Perceptron for the predictive modelling because their research shows efficiency and effectiveness in financial market prediction especially dealing with time series data.

Moving on to continuous predictive modelling, the paper by (Danqi, Charles, Valerie, & Xiaoyan, 2021) implemented regression and investigate relationships between variables of different specifications. Another paper by (Bradleya, Hanousek, Jaame, & ZiCheng, 2021) performed their research by building different regression models with different sets of data as well as different predictor/response variables. As a result, they obtained evidence showing that user comments contain useful information for the prediction of GameStop’s returns one-month ahead. However, the accuracy and predictions they had become invalid after the frenzy. Both papers used coefficient to prove the significance relationship of variables to their respective response variable. On the other hand, the paper by (Anand & Pathak, 2021) also applied regression with heteroskedasticity and auto correlation consistent (HAC) errors while GameStop’s daily return as dependent variable. Carvajal used Pooled OLS Regression as his preferred modelling method with using R-squared as the metrics to measure model performance. The end result of his model did not perform well with a potential reason of unsuitable features was proposed (Carvajal, 2021).

2.4.3 Other Analytics on GameStop

Research done by (Tolga & Melo, 2021) investigates on how successful the investment strategy of the community in “r/WallStreetBets” is. They did their research by combining transaction related words such as “buy” and “sell” and compare which has higher counts in this context. In the end, they obtained result showing that an investment strategy that follows the buy signal will be more successful in the long term with proactive buy signals outperforming reactive signals. At the same time, (Carvajal, 2021) was seen applying heatmap in his work to show the correlation between variables. What was learned from their paper is that their approach could be an inspiration and reference for diagnostic analytics to see how they are performing.

(Danqi, Charles, Valerie, & Xiaoyan, 2021) also came up with different analytics types for their research. They perform network analytics by obtaining the connectedness of comments by dividing actual connection with neighbour, over possible number of connections with neighbours. They obtained result that shows high connectedness is one of the features can significantly predict higher next-day return.

2.5 Conclusion and Remarks

Table 2: Conclusion of literature review organised following the methodology adopted

	Section	**Papers**	**Contribution**
Business Understanding	Background understanding on business case	(Kerckhoven & O'Dubhghaill, 2021)	Showed that one of the reasons for this frenzy is due to the Covid-19 pandemic
		(Burnette, 2021)	There are 3 main factors: 1. Change of demographics in retail investors 2. Presence of technology and internet 3. Social media
Data understanding	Data Mining Tool	(Danqi, Charles, Valerie, & Xiaoyan, 2021) ***AND*** (Anand & Pathak, 2021)	A common tool for data scraping from Reddit called "Push Shift API" could be used
	Data Exploration	(Danqi, Charles, Valerie, & Xiaoyan, 2021)	Found out that the returned comments contain emojis, and they considered of that too.
	Data quality	(Belo, Erker, & Koehler, 2021)	Poor data quality where a lot of post is non-financial related(noise). Example: most amazon posts are their complaints on products/services instead of financial
Data preparation	Data Cleaning	(Belo, Erker, & Koehler, 2021) ***AND*** (Joshi, Bharati, & Rao, 2016) ***AND*** (Danqi, Charles, Valerie, & Xiaoyan, 2021)	Importance and necessity of an enhanced dictionary with financial words for sentiment analytics by referencing to McDonald's work for inspiration.

		AND (McDonald & Loughram, 2015)	
	Data Cleaning	(Carvajal, 2021)	Proper feature selection is required. Social media comments and upvotes is not enough for analysis.
	Data Construction	(Belo, Erker, & Koehler, 2021) ***AND*** (Carvajal, 2021) ***AND*** (Danqi, Charles, Valerie, & Xiaoyan, 2021)	Computing the volumes and counts through the number of comments or posts
Modelling	Predictive Modelling - Regression	(Bradleya, Hanousek, Jaame, & ZiCheng, 2021) ***AND*** (Carvajal, 2021)	Implemented regression for their predictive models
	Predictive Classification Modelling – LSTM, Random Forest, and SVM	(Mehta, Pandya, & Koetcha, 2021) ***AND*** (Joshi, Bharati, & Rao, 2016)	Showed that Long Short-Term Memory (LSTM) is the best modelling technique for prediction when compared with Naïve Bayes, Linear Regression, Maximum entropy, Decision Tree, and Linear SVC Classifier in their research. The 2nd paper by Joshi, Bharati and Rao showed Random Forest(RF) and SVM performing better than Naïve Bayes. Conclusion: LSTM -> RF/SVM -> Linear SVC classifier -> maximum entropy -> naïve bayes

	Predictive Classification Modelling - MLP	(Nti, Adekoya, & Weyori, 2020)	Adopted Multi-Layer Perceptron because it is efficient and effective according to their studies, especially suitable for time series forecasting
Evaluation	Regression Modelling metrics	(Carvajal, 2021)	This paper used R-squared as the metrics to determine accuracy
	Classification Modelling metrics	(Joshi, Bharati, & Rao, 2016) ***AND*** (Mehta, Pandya, & Koetcha, 2021)	The authors used ROC Area, Precision, and Recall assessing the model performance. Mehta, Pandya, and Koetcha also used accuracy to compare the performance of classification techniques
	MLP Classification Modelling metrics	(Nti, Adekoya, & Weyori, 2020)	Performance metrics such as specificity, sensitivity, RMSE, MAPE, and accuracy were used

Based on the research done and literature review performed, many interesting insights were obtained. First and foremost, in business understanding, it is seen that potential factors lead to the happening of short squeeze on GameStop could be due to the pandemic (Kerckhoven & O'Dubhghaill, 2021) which encourages younger audiences to engage into new trends which in turn explains the demographic change in retail investors. The presence of technology and internet stimulate the ease of networking and information gathering, while social media is the medium that promotes the spread of such information to everyone who is connected to the web. Then moving on to data understanding, the papers referenced had a commonality whereby all of them used a tool to mine data from Reddit which is called the “Push Shift API”. This tool is developed by Reddit for developers to be used on Python for data scraping and analytics. With the main tool to mine text data identified, the literature review also showed that there are some papers took into consideration of emojis and calculated their sentiment manually. This was not foreseen beforehand and could be considered in the work to be performed later. If the emojis are not considered, one can see it as noisy data which leads to the next point. Belo, Erker and Koehler showed in their paper that data cleaning should taken seriously and carefully as the texts or comments returned might not be finance related and should be removed if it is not under the consideration of the topic.

In the data preparation stage, (Belo, Erker, & Koehler, 2021) , (Joshi, Bharati, & Rao, 2016) , (Danqi, Charles, Valerie, & Xiaoyan, 2021) , and (McDonald & Loughram, 2015) suggested an enhanced dictionary of positive and negative words or words is important to increase accuracy of model with a reference to research done by (McDonald & Loughram, 2015). This indicates that the implementation of this enhanced dictionary is a must to ensure a higher accuracy return. (Carvajal, 2021) suggests that proper feature selection is required as he noticed that social media comment and upvote counts are insufficient for good prediction modelling. Not only that, but a number of papers also performed data construction by counting the volume of comments and upvotes of posts and created new variables and considered them. Generally, in the data preparation stage, the most important insight and lesson obtained was the necessity of creation of a new dictionary of financial words while emphasizing in data cleaning to keep relevant contents and remove unnecessary comments to increase the overall performance for modelling.

Moving on, it was seen that most of the papers performed some sort of predictive modelling with different methods. The most common was liner regression for continuous prediction while the common method for categorical prediction was LSTM, Random Forest, and even MLP. There is no one best model to be implemented based on the research but it depends on the type of variable we are trying to predict. Obviously if we are predicting continuous data which is the stock price, regression should be implemented. However, if we are predicting categorical data, LSTM is proven to be the top performing methods while if we are factoring into time as one of the predictor variables, MLP will be a better approach. It all goes back to the question or hypothesis set, while keeping in mind that there is no one perfect model. Lastly on model evaluation, the most common metrics to measure performance for regression models are R-squared. For classification models, many metrics could be applied such as precision, recall, and accuracy to assess the performance of model. Finally, for MLP, specificity, sensitivity, RMSE, and accuracy was seen as the metrics for model performance.

Throughout the literature review, the thoughts of many researchers' work were organized and extracted to fit back to the question. Insightful results were obtained and had provided a clearer picture of the future research work. However, it is noticed that none of them used the data obtained from social media and perform clustering. This raises an interest that the future work could potentially include clustering of reddit users based on their sentiment and mood to see the general types of people that are active in the subreddit community. The result of this clustering could potentially help in extending the work by (Hasso, Muller, Pelster, & Warkulat, 2021) where they investigate who participated in the GameStop frenzy while providing more insights and clarity to the research in such topics.

3. Research methodology

For this project, the research methodology will be applied is the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. The CIRSP-DM methodology and SEMMA methodology had been in a hot debate to be the best methodology for data science and data mining project methodologies since last 2 decade. (Palacios, Toledo, Hernandez, & Navarro, 2017) performed a comparison between CRISP-DM and SEMMA, which they concluded that CRISP-DM is a methodology with detailed phases, tasks and activities while the SEMMA is more suited to work with the SAS Enterprise Miner tool. Based on their studies and own evaluation, CRISP-DM was selected due to the nature of this project is covering a wide scope and work which requires detailed planning and analysis of work to ensure the success of this project.

3.1 Research Objectives

As proposed in section 1 of project aim and objectives, the following objectives were identified to be achieved to better achieve the aim of this research project. In this section, we will be discussing on the plans to achieve these objectives accordingly. By following the CRISP-DM methodology, the plans will be planned and suited with the priority of achieving said objectives while providing sufficient details in the work.

Objective 1: To investigate the sentimental values of the community about the short squeeze alongside with the keywords by generating a word cloud

Objective 2: To build a clustering model and clusters the types of users on the reddit subcommunity based on their comments

Objective 3: To build a predictive model and predict the change of stock price of GameStop in the future

For research objective 1, “To investigate the sentimental values of the community about the short squeeze alongside with the keywords by generating a word cloud”, the rationale behind this objective is to identify the overall sentiment polarity of the community in “r/WallStreetBets” to see how they feel across time while obtaining the few most mention words by them and visualize in a word cloud. With this, we are able to put ourselves in their perspective and better understand their opinions and view on this frenzy. With this objective in mind, after the process of scraping and cleaning of data, the main task to be done is to obtain the mean value of sentiment polarity of each comment within each month and plot out together with the price movement of GameStop’s share price. As for the word cloud, a counter will be created to calculate the frequency of occurrence of each word and generate a list of top mentioned words to be inserted into the generation of word cloud.

For research objective 2, “To build a clustering model and clusters the types of users on the reddit subcommunity based on their comments”, the rationale behind this objective is to identify there are how many types of people within the community and how the behave. It is believed that not all of them within the community joined the frenzy by holding shares, that some of them are they to speculate and observe the discussions, while some are active in discussion. Some may be thinking optimistically, while there might be some thinking pessimistically. This objective is to cluster them and observe there are how many groups of people and how they behave accordingly in the community. The research on section 2 literature review showed that no papers had performed clustering based on text data which is what makes this objective significant and special. Thus, to perform analytics of clustering, k-means clustering will be used to build the cluster model with features such as sentiment, subjectivity, tones, moods, word counts and more.

For objective 3, is to build predictive model for the prediction in change of GameStop’s stock price. The reason for the building of this prediction model is a proof of work and providing evidence of the relationship between the comments and the price of stock. Obviously, there is no way that comments can affect the price of stock directly, however, the sentiment, tones, and moods expressed within the comment section might affect the decision makings of an individual towards their action on the stock. Thus, this prediction model serves to investigate the relationship between the features extracted from the text on Reddit, against the change of stock price. Based on sufficient research done in Literature Review, continuous prediction model will be built using Regression with R-squared (R^2) as the metrics to measure model fit while accuracy as the model accuracy.

Table 3: Summary of objectives in the view of types of analytics

Types of analytics	**Explanation**	**Plans / works / Objectives**
Descriptive analytics	What happened	- **Objective 1**: Generation of time series plot with change of sentiment polarity in months and change of stock price as y-axis - **Objective 1**: Word cloud generation on most frequent words
Exploratory analytics	Why did it happen	- **Objective 2**: Clustering of types of people in the community
Predictive analytics	What is likely to happen in the future	- **Objective 3**: building an accurate regression model
Prescriptive analytics	What is the best action to take	- **Objective 3**: analysis on the result on the prediction model

Generally, the research objective and its work could be explained from the view of analytics as seen in table 3. Descriptive analytics which investigate "what happened?", was answered where a time series plot of sentiment polarity change each month and the price of GameStop's stock was plotted as the y-axis while time being the x-axis to visualize what is going on. We can obtain clear and visualizable results indicating the sentiment, mood, tone, change of the community within the community to understand how the situation and the affect to their psychology is. Not only that, the word cloud generation in objective 1 will be able to show the most frequent used words to really see what they really care during the GameStop frenzy. In diagnostic analytics that is answering "why did this frenzy happened?", was answered by many papers and online sources available readily. However, another approach of perform clustering analytics was not seen adopted by many papers in the understanding of this event which this gap showed the significance of performing exploratory analytics. The clustering of types of people in the community in objective 2 could better help explain why this frenzy happened by clustering people and see how they behave in terms of commenting on posts which may explain their behaviour may indirectly promote the occurrence of this event. Then, predictive analytics which asks, "what is likely to happen in the future?", was answered through the building of predictive regression model. This will tell which of the features extracted from texts in the comments is significant enough to affect the price of GameStop and provide guidance and prediction of the change of the price in the future. Lastly at prescriptive analytics which asks, “what is the best action to take?”, will be answered by the analysis of objective 3 which is the predictive model. The significance of any features in the predictive model can tell the adopters of this model where the rise or fall of the feature could lead to a respective rise or fall in the price and suggests whether a buy or sell action should be taken.

3.2 Project Planning (CRISP-DM)

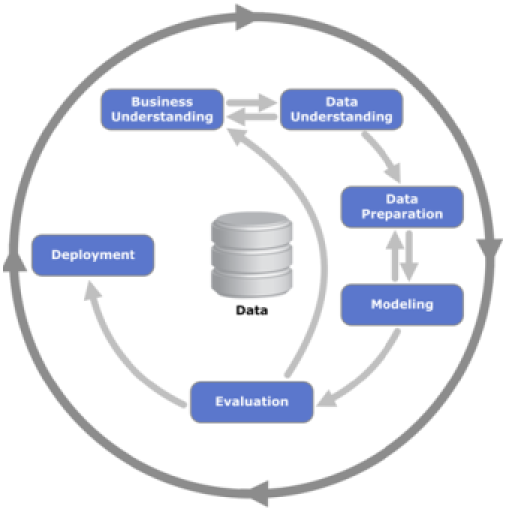


Figure 6: Image representation of the CRISP-DM methodology (Rodrigues, 2020)

Within this section, the detailed task and activities in each phase of the CRISP-DM methodology will be discussed with sufficient justification to provide clarity in the works. The first phase is "Business Understanding" talks about understanding the objectives and requirements of this project which is the story of the GameStop frenzy and why it is significant to carry out research. Next, the second phase is "Data Understanding". In this phase, is a support and addition to the foundation of the previous phase where we start to identify, collect, and perform some analysis on the data sets to better understand the GameStop frenzy. Moving on, the third phase is "Data Preparation". This phase focuses on the preparation of the final dataset based on our understanding from the previous phase where the tasks could be seen doing in this phase are data cleaning, library building, tokenization, classification and more. Next, the fourth phase which is "modelling". The final dataset from phase 3 will be implemented in this phase to create our predictive model. In the "Evaluation" stage, we will be looking at the performance of the model based on different metrics to determine whether is it good enough or it requires a revert back to phase 3 for data cleaning or feature extraction. Once the model is accurate enough and contains insights, we finally reach phase 6 which is the deployment of models and analytics for explanation. The key activity in the last stage is to explain what happened throughout our modelling and provide justification to such results.

3.1.1 Phase 1: Business Understanding

The first and most important phase of the CRISP-DM methodology starts with a business understanding. The purpose of this phase is to provide clarity into the understanding of the objectives and requirements of the project. They key activities that should be done in this phase should include a definition in research objective, assessing the situation, determining data mining goals and produce a project plan. The activities to be done in the 1st phase of the methodology is done within this planning document with sufficient proof of work. The breakdown of work done according to key activities could be seen and explained next.

1. Definition in research objective

In the first section of this planning document, research and understanding about the frenzy happened on GameStop had been performed. It occurs at the last week of January 2021 where a sudden and sharp increase of price in GameStop's share price to an all-time high of $483 per share from an average of $10 to $20 in the 4th quarter of 2020. This surge in price was seen as a work of the participation of the community in a subreddit group called "r/WallStreetBets" as an action against the hedge funds that is selling the shares of GameStop heavily. Papers had showed that the occurrence of this event could be due to the arrival of the Covid-19 pandemic, the popularity of online stock brokerages like RobinHood, and eToro, as well as the involvement of social media. With works showing such result, it leads to the formation of the main topic in this research topic which is "Text Analytics on the GameStop frenzy". With this main topic in mind, objectives were identified and to be carried out using text data with the analytics way. The first objective of investigating the sentimental value within the community was identified to understand how the community really feel about this frenzy. Are they really optimistic and believed with strong faith or they are just hopping on into the wave and trend in hopes for some profit is the main question to be answered in this objective. The result of this objective could provide insights into understanding of the community's psychology and perhaps explain their behaviour. The 2nd objective of performing a cluster on the types of people in the community was identified with the idea of further exploring the insights within the online community. Texts in the comment section could possibly help show the pattern of different types of people showing different kind of comments. For example, someone who is pessimistic and would like to exit their position in this frenzy could leave a sad mood and sad tone in the comment in perhaps of convincing others that this is the end. This shows the significance of clustering people through text. The 3rd objective of building predictive model was identified as another objective is to prove the power of text analytics in discovering insights as an indirect cause of a real-world event. There are a lot of features could be extracted from text which could be used as variables in not only the clustering but also predicting. Many papers showed work proving that text in comment sections can build accurate predictive models in the movement of share price.

With this, the research objective in this research paper is aimed to utilise text data and provide insights about the comment section of a subcommunity in Reddit called "r/WallStreetBets". The objectives were planned in mind to not be too ambitious while being able to fill in some gaps between research papers and provide insights from another point of view.

2. Assessing the situation

Assessing the situation describes the availability of resources, requirements, and risks about this research project. It is important to understand the current situation about the project to tweak the works and detail for an accurate and significant project. In this research topics, the data will be obtained and mined from online resources and there is no need to collect any first-hand data manually. This is justified where the nature of this research project was the investigation on comments towards the effect of share price instead of people's opinion where first-hand data collection will benefit from. The risk of this project was found where there is a possibility of having a poor performing model due to insufficient data, irrelevant features and more. These risks will directly impact the outcome and significance of this research topic. Thus, risk-reduction measures are proposed and included within the working document. For example, a longer time span of text data should be mined to ensure a sufficient large data which is up to 1 year. Not only that, but the features will also be extended to include verb and adverb density, noun density, to ensure that there are sufficient interesting features to be included in the model. Lastly, as reviewed and proposed, a dictionary enhanced with financial words will also be included to ensure better accuracy of sentiment polarity identification to further reduce risk of having a poor model.

3. Determining data mining goals

After successful definition in research objective, data mining goal should also be determined to see how a successful project looks like from a data mining perspective. From a general view, the big goal from the data mining perspective is to be able to produce high quality codes based on the written plans. Viewing it from the details, is to be able to mine all the relevant social media text data successfully within the proposed targeted dates and ensure the dataset is sufficiently large. Not only that, but the other detailed data mining goal should also be able to build an optimal dictionary of positive and negative words for successful sentiment analytics. Lastly, the data mining goal should also include an accurate predictive model.

4. Produce a project plan.

The final activity to be performed in the first phase of CRISP-DM methodology is about the production of a project plan. This working document should serve as the project plan with details about the research.

Table 4: Summary of activities and tasks to be done in the business understanding phase

Activities	Tasks
Definition in research objective	- Defined a list of objectives - Provided justification of the proposing the list of objectives
Assessing the situation	- Data will be obtained and mined online - No first-hand data collection required - Proposed 3 risk-reduction measures to be implemented
Determining data mining goals	- To produce high quality codes - To be able to mine all data within the proposed targeted date - Building an optimal dictionary of words
Produce a project plan	- Produce a detailed working document

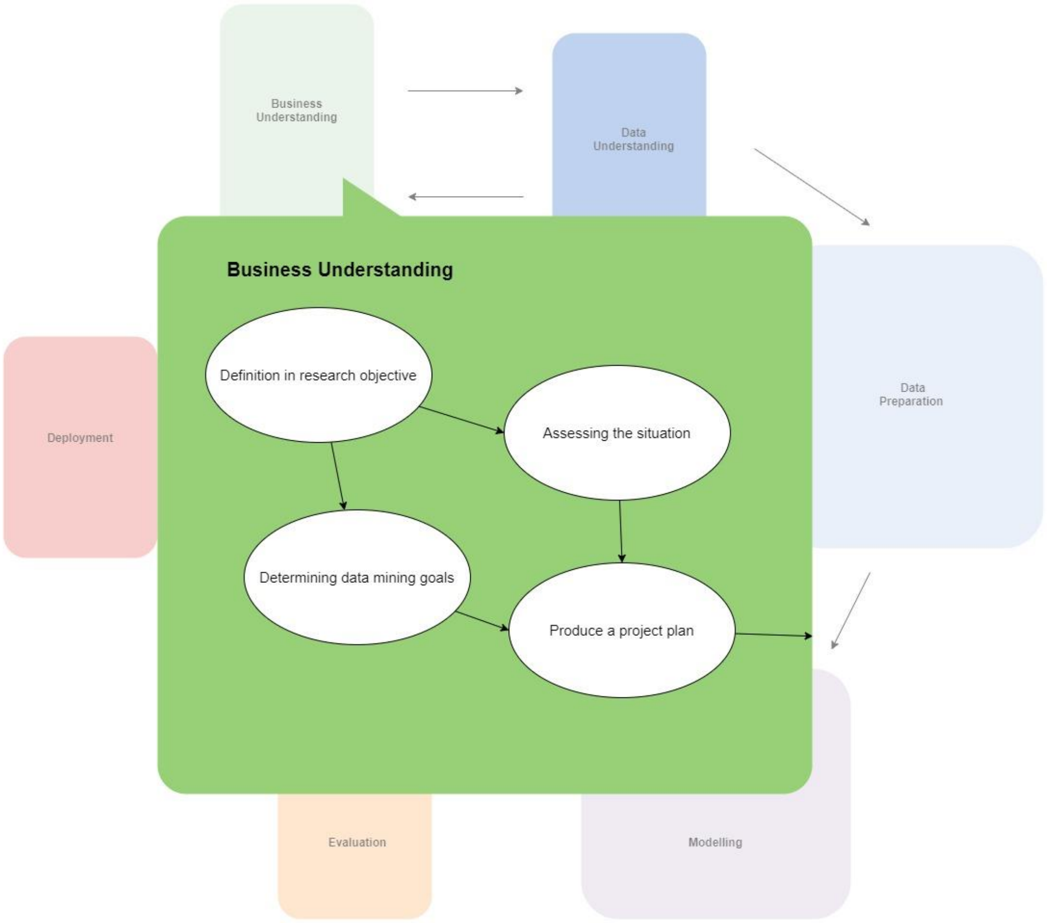


Figure 7: Flow of activities to be done in the Business Understanding phase

Table 4 shows the summary of activities and detailed tasks to be done in the first phase of the CRISP-DM methodology adjusted to this research project with figure 7 showing the flow of them. In summary, the findings and advice from literature review were applied in the activities of this research topic and was planned with a series of tasks to which was proven to be working with sufficient risk control.

3.1.2 Phase 2: Data Understanding

After the business understanding phase, the next phase to be done is data understanding. The purpose of this phase is to provide additional support to the foundation of business understanding. The main idea of this phase in supporting business understanding is to understand the data we are mining and all the details about it. Without sufficient data understanding, it is impossible to proceed for analysis because there is a very high risk of project failure due to insufficient or incompatible data. The core of this section is to avoid the concept of "garbage in, garbage out" at the most. Thus, data understanding not only is to understand it, but also identifying the features as well as potential risks and costs. The main activities to be covered in this phase are collecting initial data, describe data, explore data, and verify data quality.

1. Collecting initial data

There are 2 main data sets in this research project, which one of it is the text data from the comment section of the subreddit page "r/WallStreetBets" from Reddit while the other is the history prices of GameStop. The main programming language and software to be used for data collection is Python due to the wide availability of API and tools. For text data collection, a python tool called "PushShift API" will be used to mine data from Reddit with an installation of the "praw" tool is required prior to that. This tool was developed by Reddit themselves and was present in works of many research papers as seen in the literature review. The comments will be mined from the posts between October 2020 to October 2021 by stating the before date and after date inside the "search_comments" function. With that, the text data from Reddit is obtained successfully and ready for the next phase. Another set of data will be required for the analytics which is the history price of GameStop. For this, the data could be downloaded from the Nasdaq's official website which is in Comma Separated Value (CSV) format in this link. A csv file will be downloaded containing columns of data that records the historical value of open, close price of GameStop. Generally, the timeline of text and stock data to be collected is from 1st October 2020 till 1st October 2021. This timeline is chosen because the frenzy occurs at end of January 2021 which is the midpoint of this timeline. 1 year long of data will be obtained is to also ensure that the dataset is sufficiently large for proper training and validation. Now the datasets are obtained, next is the building of an advanced dictionary of positive and negative words. The original dictionary will be obtained from a python tool called "SentimentIntensityAnalyzer" by NLTK because it is commonly used and specially designed to handle social media words like emojis, slangs, and emoticons. The dictionary is then expanded using the words list by (McDonald & Loughram, 2015) to enhance and better fit with financial words for accurate analysis.

2. Describing data

After mining and obtaining initial data, we can observe a few points of the returned data. First and foremost, it was seen from table 5 that the returned data after the mining using "PushShift API" returns many data about the comments that consists of 37 columns. Some columns of the data are related to the authors such as author name, author flair type, awards, while the rest are related to the data of the comment block such as the link, score, link_id etc.

Table 5: Raw data returned from PushShift API

Column Name	**Python Type**	**Data Type**	**Description / Definition**
all_awardings	Object	Key-Value List	All awards the user were awarded in Reddit
Associated_award	Float64	String	Awards the author had associated with the subgroup
author	Object	String	Reddit username of the author
Author_flair_background_color	Float64	String	Hexadecimal color code of the author's
Author_flair_css_class	Float64	String	The css class of the author's flair(tag)
Author_flair_richtext	Object	Key-Value List	A dictionary of flairs(tags) the author uses
Author_flair_template_id	Float64	String	The template ID of the author's flair
Author_flair_text	Float64	String	The text of the flair itself
Author_flair_text_color	Object	String	The color of the author's flair. Light/Dark/None only
Author_flair_type	Object	String	The type of the author's flair. Richtext / text only
Author_fullname	Object	String	Reddit ID of the author
Author_patreon_flair	Boolean	Boolean	Whether the author has a pateron flair. True/false only

Author_premium	Boolean	Boolean	Whether the author had subscribed to reddit premium
awarders	Object	Key-Value List	A list of awarders to the author
Body	Object	String	The content or text of the comment
Collapsed_because_crowd_control	Float64	String	Whether the feature crowd control feature was enabled
Comment_type	Float64	String	Types of comment
Created_utc	Object	String	Datetime of the comment created
Gildings	Object	Key-Value List	Gildings the author of post received
Id	Object	String	ID of comment
Is_submitter	Boolean	Boolean	Whether the commenter is the author of the parent post
Link_id	Object	String	An ID which is part of the permalink
Locked	Boolean	Boolean	Whether the thread was locked
No_follow	Boolean	Boolean	Whether the author can be followed on reddit
Parent_id	Object	String	ID of the parent comment
Permalink	Object	String	URL or link to post
Retrieved_on	Object	String	Datetime when comment was retrieved through pushshift API
Score	Int64	Continuous integer	Score(likes) of the comment
Send_replies	Boolean	Boolean	Whether the comment author accepts replies
Stickied	Boolean	Boolean	Whether it is a ‘sticky post’

Subreddit	Object	String	The subreddit where comment was posted
Subreddit_id	Object	String	The ID of the subreddit
Top_awarded_type	Object	String	Top given award types
Total_awards_received	Int64	Continuous integer	Number of total awards received
Treatment_tags	Object	Key-Value List	Types of treatment tags
Distinguished	Object	String	Whether the author is the 'moderator'
Author_cakeday	Boolean	Boolean	Whether the comment is the date where the user originally signed up on Reddit

Table 6: Raw financial data downloaded from Nasdaq

Column Name	**Data Type**	**Description / Definition**
Date	Date	The date of the financial data in month/day/year format
Close/last	Float	The closing price of the share that day
Volume	Integer	The number of shares traded that day
Open	Float	The opening price of the share that day
High	Float	The highest price of the share that day
Low	Float	The lowest price of the share that day

Descriptions and identification of data types about the raw data returned right after mining was shown in table 5. It is seen that not all 37 columns of them are useful but rather a few only. It was also seen that most of the data were strings, with a minority on key-value list(dictionary), Boolean and integers. At the same time, table 6 shows the financial data set downloaded from Nasdaq had 6 columns of data that is related to the price of share at the day. All of them are useful but further processing is required to make all the data more meaningful.

3. Explore data

As shown in table 5 and table 6 are the dataset that this paper will be dealing with in focus. All the data has their own meaning and purpose that differs a lot among them. Upon further exploration at table 7, it was noticed that the raw dataset returned from Reddit has 37 columns with various data.

14 variables about the author of that comment, 1 for the body of the comment itself, with the remaining 22 variables are about the comment. The variables about the author are mostly on the display styles like awards, CSS class, text, and text colour, without really including other information on author such as start date. Thus, upon consideration, all the variables related to the author of that comment will be removed to not add weight and complexity to analytics except for the author's Reddit ID. Moving on, the body of the comment itself will definitely be included because it is the main content we are looking for upon data mining. For the remaining 22 variables that contains data about the comment, are various settings and features that does not bring significance in text analytics. Thus, most of them will be removed except for a few such as ID, datetime, and score because they are used as identification purpose, datetime footprint, and shows popularity in the comment respectively. Table 7 records the final decision of variable keeping with justification alongside.

Next, the raw financial data downloaded from Nasdaq as shown in table 6 consists of 6 columns which all is about the stock movement of GameStop for the day. All the variables were meaningful and carries significance for analytics later this all of them will be remained. Table 8 records the final decision of variable keeping for financial data with sufficient justification and purpose.

Table 7: Variable keeping/dropping decision upon exploration of Reddit dataset

Column Name	Description / Definition	Keep?	Justification
All awardings	All awards the user were awarded in Reddit	No	Awardings does not bring significance
Associated award	Awards the author had associated with the subgroup	No	Associated awards does not bring significance
author	Reddit username of the author	No	Author ID was used for identification already
Author flair background color	Hexadecimal color code of the author's	No	Styles does not bring significance
Author flair css class	The css class of the author's flair(tag)	No	Styles does not bring significance
Author flair richtext	A dictionary of flairs(tags) the author uses	No	Styles does not bring significance
Author flair template id	The template ID of the author's flair	No	Styles does not bring significance

Author flair text	The text of the flair itself	No	Styles does not bring significance
Author flair text color	The color of the author's flair. Light/Dark/None only	No	Styles does not bring significance
Author flair type	The type of the author's flair. Richtext / text only	No	Styles does not bring significance
Author fullname	Reddit ID of the author	**YES**	Used for identification of author
Author patreon flair	Whether the author has a pateron flair. True/false only	No	Patreon identification does not bring significance
Author premium	Whether the author had subscribed to reddit premium	No	Premium identification does not bring significance
awarders	A list of awarders to the author	No	Awarders does not bring significance
Body	The content or text of the comment	**YES**	The main content of text data
Collapsed because crowd control	Whether the feature crowd control feature was enabled	No	Comment settings does not bring significance
Comment type	Types of comment	No	Comment type does not bring significance
Created utc	Datetime of the comment created	**YES**	To know the datetime and perform comparison
Gildings	Gildings the author of post received	No	Gildings does not bring significance
Id	ID of comment	**YES**	Used for identification of comment
Is submitter	Whether the commenter is the author of the parent post	No	Submitter status does not bring significance
Link id	An ID which is part of the permalink	No	Permalink ID does not bring significance
Locked	Whether the thread was locked	No	Lock status does not bring significance
No follow	Whether the author can be followed on reddit	No	Follow status does not bring significance
Parent id	ID of the parent comment	No	Parent-children relationship is not under area of analytics

Permalink	URL or link to post	No	Link does not bring significance
Retrieved on	Datetime when comment was retrieved through pushshift API	No	The datetime is common across all, does not bring significance
Score	Score(likes) of the comment	**YES**	Number of likes could show popularity/agreement on the comment
Send replies	Whether the comment author accepts replies	No	Reply status does not bring significance
Stickied	Whether it is a ‘sticky post’	No	Sticky post does not bring significance
Subreddit	The subreddit where comment was posted	No	Common across all comments, thus does not bring significance
Subreddit id	The ID of the subreddit	No	Common across all comments, thus does not bring significance
Top awarded type	Top given award types	No	Award types does not bring significance
Total awards received	Number of total awards received	No	Total awards does not bring significance
Treatment tags	Types of treatment tags	No	Treatment tags does not bring significance
Distinguished	Whether the author is the ‘moderator’	No	Does not bring significance
Author cakeday	Whether the comment is the date where the user originally signed up on Reddit	No	Does not bring significance

Table 8: Variable keeping/dropping decision upon exploration of stock dataset

Column Name	**Description / Definition**	**Keep?**	**Justification**
Date	The date of the financial data in month/day/year format	**YES**	For identification purpose
Close/last	The closing price of the share that day	**YES**	To calculate price difference later of the day
Volume	The number of shares traded that day	**YES**	To identify the number of trades at the day

Open	The opening price of the share that day	**YES**	To calculate price difference later of the day
High	The highest price of the share that day	**YES**	To identify the fluctuation of price in the day
Low	The lowest price of the share that day	**YES**	To identify the fluctuation of price in the day

4. Verify data quality

Table 9: Data quality of Reddit dataset

Variable Name	**Example**	**Data Quality**
Author_fullname	t2_4vtqk2e5	There is a pattern that it starts with "t2_". No data quality issue as it is for verification purpose only
Body	Local Gamestop ran out of gift cards. FAIL. Puts	It is in a cleaned string format. No data quality issues
Created_UTC	1608071895	It is an EPOCH timestamp in UTC format. No major data quality issues but may require conversion
Id	gfyysa3	No data quality issue as it is for verification purpose only
Score	15	It is a cleaned integer. No data quality issues

Table 10: Data quality of stock dataset

Column Name	**Example**	**Data Quality**
Date	9/28/2021	It is in months/day/year format. No major data quality issues
Close/last	$178.60	It has decimal which is in float format. No major data quality issue other than removal of dollar ($) sign
Volume	1770493	No data quality issue as it is straightforward
Open	$188	It has decimal which is in float format. No major data quality issue other than removal of dollar ($) sign
High	$190.81	It has decimal which is in float format. No major data quality issue other than removal of dollar ($) sign
Low	$178	It has decimal which is in float format. No major data quality issue other than removal of dollar ($) sign

As shown in table 9, the shortlisted Reddit dataset shows data in different format. Generally, the data are clean and requires minimal pre-processing work except for the date and time. The variable "Created_UTC" requires a conversion of the EPOCH timestamp to a human readable format or even a separation into 2 independent variables of date, and time. Table 10 at the same time also showed clean data overall except removal of the dollar ($) sign across some variables to make it a proper float variable. Other than that, a common date time format should be applied to both datasets so that they are unified and of the same position to ensure higher data quality.

Table 11: Summary of activities and tasks to be done in the Data Understanding phase

Activities	**Tasks**
Collecting initial data	- Text data will be mined from Reddit using python and a tool called "PushShift API" - Stock data will be downloaded from Nasdaq's official website - Positive and negative word lists will utilize a python tool called "SentimentIntensityAnalyzer" and enhanced with McDonald's dictionary
Describing data	- Created 2 tables that explains the type of data and their description for text data and stock data respectively
Explore data	- Further explore the variables with their meanings and provide justification alongside the decision of keeping them
Verify data quality	- Verified the data quality with 2 tables for each dataset and stating the possible issues

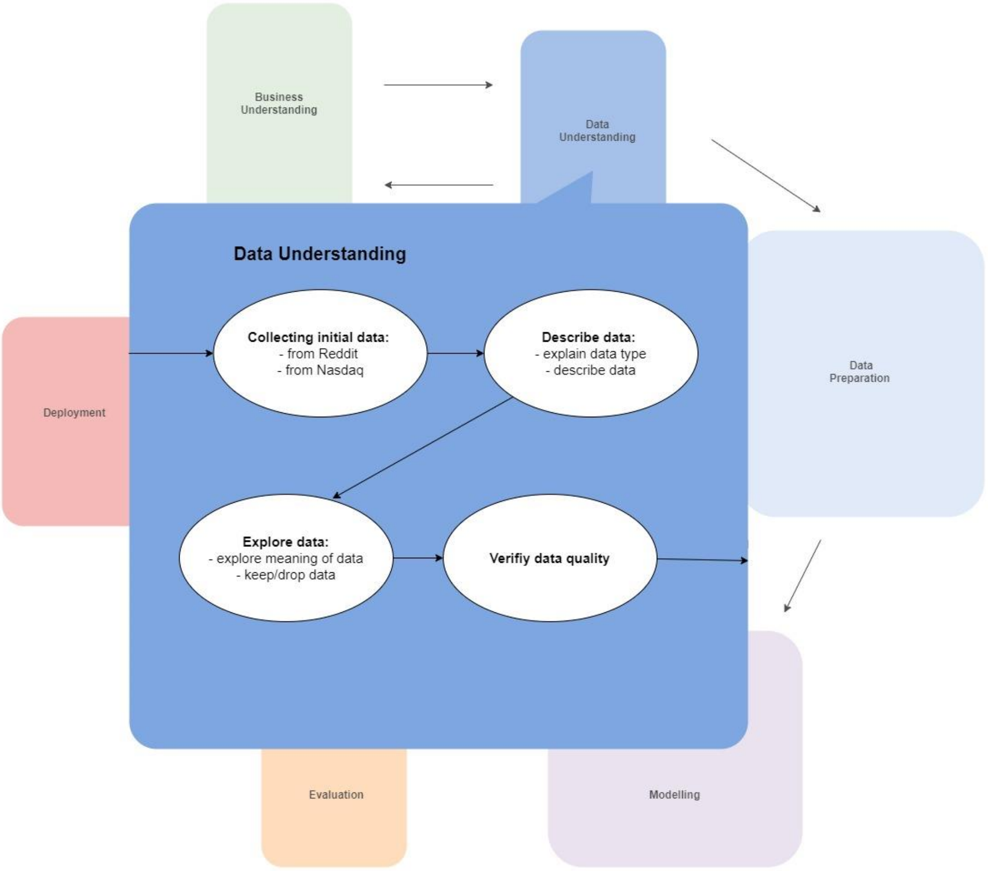


Figure 8: Flow of activities to be done in the Data Understanding phase

Table 11 shows the summary of activities and their tasks performed in the data understanding phase of the CRISP-DM methodology with figure 8 visualising the flow. In general, it is seen that the understanding and exploration towards the behaviour of data was enhanced while identified certain risks and provide strategies to overcome them.

3.1.3 Phase 3: Data Preparation

After the data understanding phase, the next stage to move to is the data preparation phase. This phase which was also known as “data munging”, brings purpose to project by preparing the final datasets into final form for modelling. With the data explored and understood thoroughly in the previous phase, this phase is primarily focused on cleaning them according to the nature of the research to ensure a relevant dataset produced for more accurate analysis. It is also a common rule of thumb that 80% of the project is focused on this stage. Thus, repetition and revisit to this phase could be seen occurring during project commencement. They key activities to be covered on this phase are data cleaning, data reformatting, data constructing, and finally data integration.

1. The text data

The first activity to be performed is data cleaning. This is the lengthiest activity that needs to be performed in great attention to detail or not there will be a high probability of being the victim of “garbage in, garbage out”. The motivation of having data cleaning is to remove incorrect and inconsistent data that could cause false conclusions and predictions.

After the text data was scraped from reddit, the unused variables will be removed following table 6. The variable was also renamed for easier reference such as the variable “author_fullname” will be renamed to “author_ID” and “id” will be renamed to “commentID”. After removal of unnecessary variables, the “created_utc” variable of the data frame will be converted from UNIX to datetime format for a unified format for dataset joining later. Once it is cleaned, the date and time were extracted and created as new variable respectively for the convenience of any data query in the future.

2. The financial data

As for the financial data, once it is read into the Python environment, the only cleaning performed is to remove the dollar sign out of 4 variables which were “Close/last”, “Open”, “High”, and “Low”. After that, these variables are converted to float type as they are integers with decimal points in nature. The financial data also undergoes a similar process where the variable “datetime” was converted to datetime format unified with the datetime variable of the text data. After that, some data construction was made to extract insights on the daily returns of the stock price. The highest price of the day was subtracted with the lowest price of the day and the result obtained was used to create a new variable named “fluctuation” to indicate the fluctuation of stock price in that day. Then, the fluctuation was divided with the lowest price of the day to obtain the percentage of fluctuation of that day named as “fluctuation_pct”. This indicates the size of the price fluctuation that day. After that, a variable called “returns” was calculated by acquiring the difference of the closing price of each day indicating the growth/fall of the stock price each day. Lastly, the percentage change of the returns was also calculated and stored in a new variable called “returns_pct”.

3. Word List

After reading in the LoughranMcDonald’s dictionary into the Python environment, the positive words are extracted and labelled “1” while the negative words are extracted and labelled with “-1”. After they are merged into a single word list, the “SentimentIntensityAnalyzer(SIA)” package will be imported from the “nltk sentiment vader” toolkit for sentiment analytics. The rationale for choosing “SentimentIntensityAnalyzer(SIA)” was mentioned previously, and it will serve as the basis for not only sentiment analytics but also word extraction and label. The dictionary of “SIA” was enhanced through the update and insertion of the word list extracted from LoughranMcDonald’s dictionary.

4. Master dataset

Once the text data, financial data, and the dictionary was prepared, a final dataset will be built with some processing to be done next. First and foremost, the text dataset and financial dataset was combined as a big dataset by using the date as the id and key. This ensures that the financial information of the day that the comment was created was encoded together with it. After that, the text comments in this master dataset were tokenized by using the “nltk tokenize” toolkit. A list of English stop words were obtained using the “nltk corpus” toolkit and it was expanded with a list of repeating words such as “GME”, “gme”, “Gamestop”, “GameStop”, “GAMESTOP”, and “gamestop” to ensure accuracy. Then, 2 functions were created for respective usage: to remove noise and lemmatizing the sentences. With the list of stop words and functions ready, the master dataset undergoes noise removal with the list of stop words, and then undergoes the second process of lemmatization. The output is a list of cleaned tokens originated from the text of comments. After that, the sentiment of each comment will be calculated while values for imageability, valence, arousal, dominance, and subjectivity will also be obtained.

Not only that, some new features or variables will also be extracted from the original text comments. To be exact, the word densities will be extracted as new variables. They are the noun density, verb density, adverb density, and adjective density. The densities will be calculated using the tags coming after the tokenized words that was present in the lemmatization process. The purpose of introducing densities as new features is to further expand the variables for hopefully a better prediction and classification model.

In the end, the tokenized merged dataset, values for sentiment polarity, imageability, valence, arousal, dominance, subjectivity and the 4 of the word densities will be merged again to produce the final dataset ready for analytics.

Table 12: Summary of activities and tasks to be done in the Data Preparation phase

Activities	**Tasks**
Text data	- Remove unused variables as mentioned in table 6 - Convert date time variable to a unified unit - Extract date and time respectively as new feature
Financial data	- Remove dollar($) sign and convert to float type for variables - Obtained 4 new features : “fluctuation”, “fluctuation_pct”, “returns”, “returns_pct”
Word List	- Extracted positive and negative word list from LoughranMcDonald’s dictionary - Enhanced the dictionary of “SIA” with the extracted word list
Master dataset	- Merged cleaned text and financial data - Remove noise and performed lemmatization - Obtain 10 new features based on words and densities - Processed dataset and new features were merged as the final master dataset

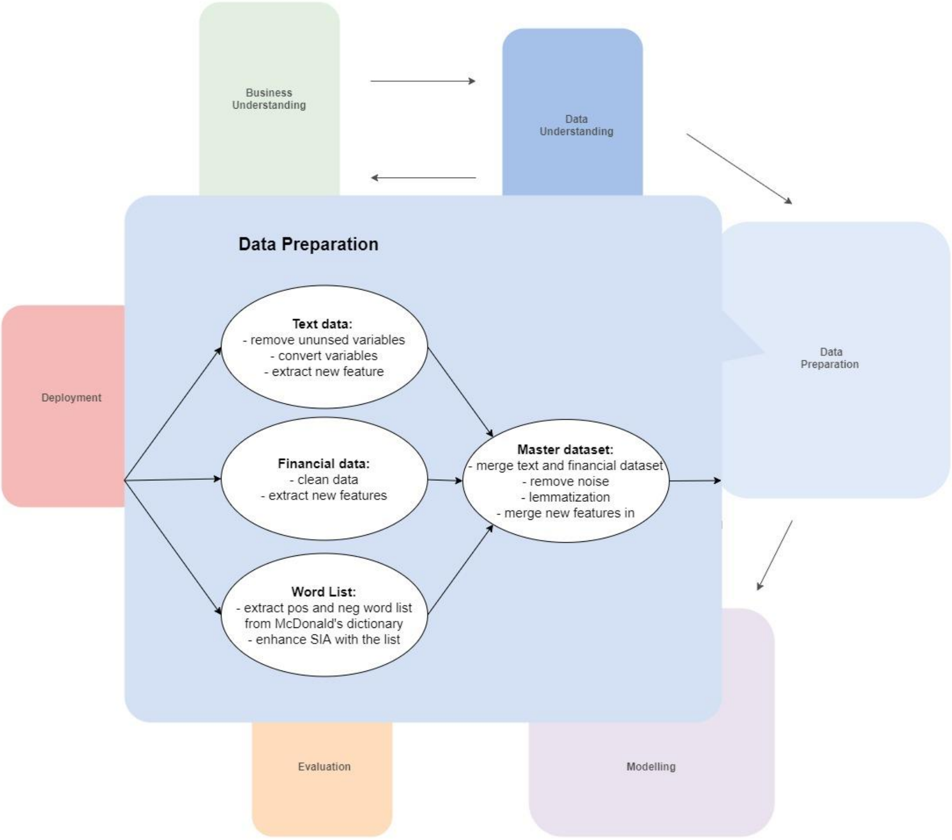


Figure 9: Flow of activities to be done in the Data Preparation phase

Table 12 shows the summary of activities and task to be performed in the data preparation phase. It is seen that a lot of data processing was performed in this stage while a lot of new features/variables were extracted throughout this phase. However, as prepared as possible, it is expected to have re-visits back to this phase during project implementation to better prepare the final data or introduce new methods that can improve the result.

3.1.4 Phase 4: Modelling

Moving on from data preparation phase is the modelling phase, where the building and assessment of models occurs on several different modelling techniques. The occurrence of this phase is after the data is prepared and is in good shape ready to harvest the insights or patterns. Different models will be applied based on the question to be answered while using different sets of data or variables instead of a unified set. They activities to be covered in this stage were selecting the modelling techniques, designing tests, building models, and finally assessing the models.

1. Selecting modelling techniques

In the analytics world, there are an abundance of modelling techniques to be chosen, but there is no go-to technique that suits all needs. Thus, this highlights the importance of selecting the most suitable modelling technique based on the availability, variables involved, data and research considerations. For example, many organizations prefer techniques which its result is easy to interpret and thus decision tree and regression will be under consideration while neural network is not. However, for a professional laboratory-controlled experiment, neural network could be the answer instead of decision tree or regression. Therefore, modelling techniques is also one of the crucial decisions that needs to be identified and made.

For cluster modelling, the technique that will be chosen for the research project is “K-means clustering”. K-means clustering is an unsupervised machine learning algorithm that identifies cluster through similarities between the values of data through the average. It was chosen because of a few reasons: it is easy to implement, has high flexibility in cluster adjustment, suitable for large dataset, easy interpretation of results, and handles numerical data well. However, the downfall is that assumptions will be considered in the operation such as the size and number of clusters. Thus, to implement k-means clustering, there will be a set of methods and assumptions experimented to obtain optimal result of the data.

For predictive modelling, as planned previously, Regression and Analysis of Covariance (ANCOVA) will be chosen for this research project. Regression is a statistical process that estimates the relationship between variables and tries to predict the effect towards the predictor variables. It was chosen because the final master dataset will contain a lot of variables which were continuous(numerical) data with the stock price being a continuous data too. This makes regression suitable for predicting the stock price based on a series of numerical data harvested from text, while identifying their effect and relationship to stock price. At the same time, ANCOVA was chosen as another predictive modelling technique due to its suitability in this research project. ANCOVA is a general linear model which is a blend of ANOVA and regression. It tells information of the dataset by considering one variable at a time, which was independent from the others. ANCOVA was selected because it can handle both continuous and categorical data with a final goal being predicting the stock price too. Decision tree was previously shortlisted as one of the modelling methods but was eliminated because it requires high requirement of device, while being very sensitive to small data changes as the number of data increases. Neural network too was under consideration but eliminated due to their structure. It has a "black box" nature which we do not know what was happening throughout the modelling process and how it leads to such results. This sways away with out motive of this research topic which is to explain what was happening. Lastly, was Bayesian network. According to (Uusitalo, 2007) one of the challenges of using Bayesian Network is the need of collecting and structuring expert knowledge. Bayesian Network was seen unsuitable for this research topic as the scope is quite big that requires professional knowledge on both words and finance to clearly understand and define the result of the model, which is why it too is eliminated. In the end, Regression and ANCOVA was chosen due to their overall suitability over the other modelling techniques.

2. Designing tests

The tests here are referring to the data to be tested. It needs to be designed in a way that the test used will determine how well the model works. Different tests will be designed based on the respective modelling techniques. However, a general test will be performed is to identify the optimal train-test data splitting to ensure best result. This will avoid the problem of overfitting and underfitting which causes imbalanced result.

For cluster modelling, another test will be carried out specifically to identify the optimal number of clusters. This is very important as it is required to insert the number of clusters prior to model building. Not only that, the "sklearn" toolkit will also be used to identify influential variables that are suitable to be implemented in the clustering model.

For predictive modelling, tests will need to be carried out to ensure that the data met the assumptions of both models. Table 13 shows the description of each assumption and the test to be carried to reach the assumptions. Generally, these assumptions should be met to ensure a good predictive model, but it is also possible to violate a few of these assumptions. Other than that, different selection method of significant variables will also be tested and experimented to obtain the best prediction. Measures such as variable removal, or a revisit to data preparation phase will be carried out to produce a better model that fits the assumptions as much as possible.

Table 13: List of assumptions for regression and ANCOVA with respective tests

Assumptions	Description	Tests
Linearity of regression	The regression relationship between response and predictor variable must be linear.	Plot regression line and observe the line
Homogeneity of error variances	The error is a random variable with conditional zero mean.	Run Levene's test and observe the significance (p-value)
Independence of error terms	The errors are unrelated to each other	Analyse significance (p-value) of covariate between variables
Normality of error terms	The residuals should be normally distributed	Plot a frequency distribution chart and compute the values of skewness and kurtosis
Homogeneity of regression slopes	The slopes of the regression lines should be parallel	Plot multiple regression lines on a single plot and observe

3. Building models

With modelling techniques identified and a list of tests done for the model, it is time to build the model themselves. The building of the models is the heart and key section of the whole methodology after many efforts invested in previous stages.

When building the clustering model, the test result of optimal data splitting ratio and number of clusters will be brought forward during this building session as settings. Then, the model will be built with influential variables shortlisted from the testing phase too to produce the result. After the result is produced, justification and labelling of each cluster will be carried out in the next stage.

For the building of the regression model, the stock price will be set as the response variable. For predictor variable, it will be shortlisted through its significance(p-value) by first fitting all variables into model with different variable selection method such as forward, backward, and stepwise. Once there is a result, the variables will be used to fit in a new model to testify and obtain its final accuracy. For the building of the ANCOVA model, similarly, the stock price will be set as the response variable with variables deemed significant through a variable selection process.

As the initial models were built and initial results were obtained, it is expected that it is not satisfying and optimal. Thus, it is also expected that effort will be seen going back to the previous phases of the CRISP-DM methodology to perform tweaking and adjustment to build more models down the road to reach the optimal results as close as possible.

4. Assessing models

Once the models were built, they will undergo reviews from a technical and business perspective in this stage. From the technical standpoint, the statistical results will be reviewed to see whether the model is satisfying. If the result is not satisfying, more tests will be performed to ensure that they met the requirements of the models and data preparation will be revised based on the indication of the model results. From the business standpoint, the results will be evaluated whether they truly explained the relationship and trend in a logical way or explained in a valuable result. The business perspective explains on the overall direction and performance of the model while the technical perspective explains whether the model is sufficient and good enough based on different metric such as p-value for variable significance, R-squared(R^2) to explain the fit of regression model, and accuracy to explain the success/failure variables predicting stock price.

Table 14: Summary of activities and task to be done in the Modelling phase

Activities	**Tasks**
Selecting modelling techniques	- Clustering model: K-means clustering - Predictive model: Regression and ANCOVA
Designing tests	- Both models: optimal data splitting ratio - Clustering model: optimal cluster numbers, variable selection - Prediction model: variable selection, assumptions
Building models	- Clustering and prediction models will be built based on the result from tests as settings
Assessing models	- Metric will be used to assess the model from technical perspective - Direction, logicality, and value will be assessed from business perspective

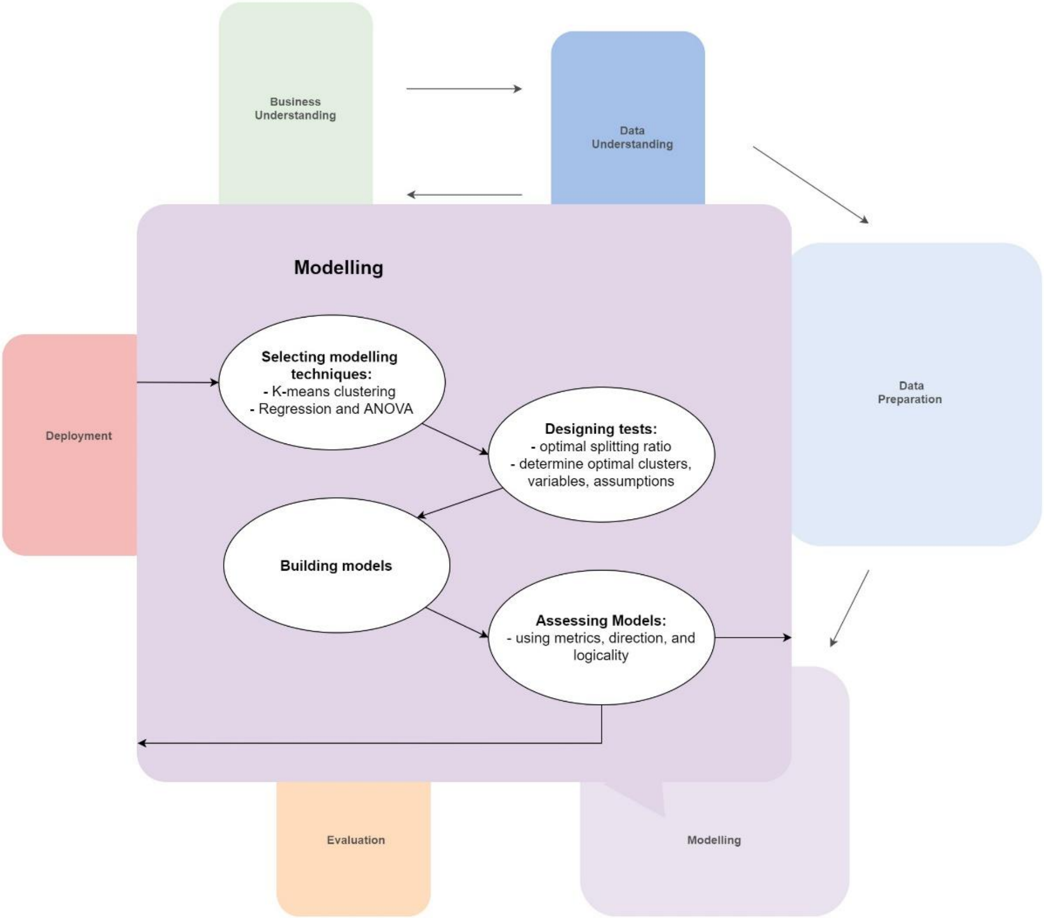


Figure 10: Flow of activities to be done in the Modelling phase

Table 14 and figure 10 shows the summary and flow of activities and tasks to be performed in the modelling phase respectively. 2 types of models with 3 different techniques were identified and to be implemented for the building of model. “K-means clustering” was chosen for clustering model, while “Regression” and “ANCOVA” were chosen for predictive model. Prior to model building, a series of tests will be carried out to experiment and identify the optimal settings for each modelling techniques. After the settings were identified, the models will be built and adjusted to produce an accurate and satisfying enough result to be explain through a technical and business perspective.

3.1.5 Phase 5: Evaluation

In the first 4 phases of the CRISP-DM methodology, the data was explored, and patterns was found, with some modelling result produced. The new question to be evaluated is: are the results good for this research project? In the 5th phase of evaluation, it is not only focused on evaluating the model, but also the process and works performed so far, with their potential for deployment and practical use. Looking on a bigger scale, the evaluation phase is to look broadly where the model and results had met the objectives and project goals while providing guidance on what to do next. The key activities that should be covered in this phase were: evaluate results, review process, and determine next steps.

1. Evaluate results

To initiate the evaluation phase, it all starts with results evaluation. All the values obtained from any process starting from data exploration till modelling is evaluated to assess whether they had met the research goals identified initially. We will be looking at any reasons why the model or results was not satisfactory for our use case.

The results obtained at the end will be summarized and interpreted to business perspective and justified whether it explains something and uncovered any insights. It is also expected to state whether the research goals defined initially were met and was satisfying enough with thorough explanation to such results. For example, hypotheses set for the regression model will be evaluated and determined whether which hypothesis will be rejected, and which will be accepted to explain such results. If expected results was not achieved, explanations and possible cause will also be identified and explained to evaluate in why such results were obtained.

2. Review process

Before we really finalise in the presentation of such methodology and results, some time will be taken to review back the roadmap of methodology and process. This important in a sense that it provides an opportunity to spot issues that might be overlooked that draws attention to flows in work. As such, there is still room and time to correct the problem before deployment to produce better results. For example, some data cleaning methods or feature extraction were overlooked previously, and the review process will come into play by providing another opportunity of redemption from mistakes. The main deliverable for this activity is a process report where the outline of review process will be recorded with mentions about the findings and highlights discovered through the process.

3. Determine next steps

The whole evaluation phase concludes with a list of recommendations for best next steps. A pre-requisite to be achieved here is that the data mining process should end already and had sufficient reviews. Thus, in this activity, the final model might be ready to deploy, or could be done better with some other process which might see improvements that can inspire a new research project. A list of possible actions should be produced with clear descriptions on each alternative action alongside with reasons of going with and against it. With a list of actions stated, a final decision should be made based on the possible actions, with string reasoning behind it.

Table 15: Summary of activities and tasks to be done in the Evaluation phase

Activities	**Tasks**
Evaluate results	- Values obtained will be assessed - Identify whether the project objectives were met
Review process	- Produce a process report that outlines the review process
Determine next steps	- Produce a list of possible actions - Identify final decision

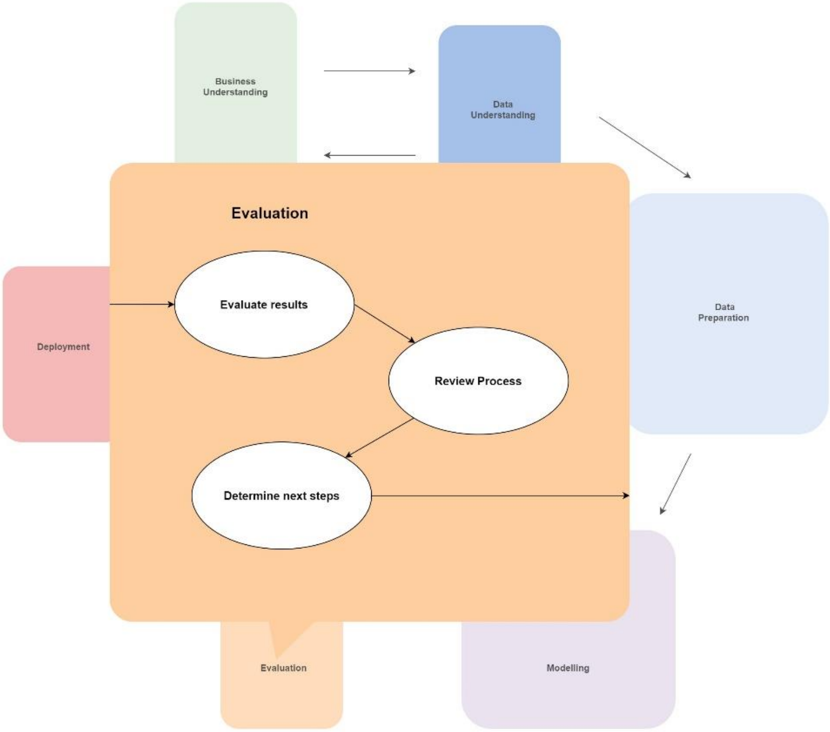


Figure 11: Flow of activities to be done in the Evaluation phase

Table 15 shows a summary of activities and tasks to be performed in the evaluation phase. Generally, the deliverables in this phase were mostly on written format on reports provide reasoning, and justification on the obtained results after a series of work throughout the research methodology. This phase will see a wrap-up of any development and analytics to bring to an end of the research.

3.1.6 Phase 6: Deployment

Finally, the last phase of the CRISP-DM methodology, which is the deployment phase. It is where the data mining process pays off and a final deliverable is presented. It does not matter how insightful the ideas were discovered if it is not deployed and presented at the end. Essentially, this phase is about a conclusion and recap of what was done and what was obtained, with why justifications all organized out in proper documentations. Thus, the key activities to be covered are planning deployment, planning monitoring and maintenance, final results reporting, and project review.

1. Planning deployment

After the models are all ready, they will require a plan to be deployed. The results will be written into formal documents as a proof of work and obtain approval and markings from panel of judges by the school. If the results returned from the university is excellent, the whole research project might be uploaded to GitHub and submitted to be included in journals for the beneficial of the public and as an effort contributing to the world of analytics.

2. Planning monitoring and maintenance

Data mining work is in a cycle, where if possible, stay actively involved with the models and integrate them into daily or periodic usage. Although the future plans are highly influenced by the results, this is a research project instead of a business use case that sees high usage after development. However, the model will be used after development to provide insights and evaluate whether a third wave will occur again. If the model performs poorly, the future plan for monitoring and maintenance will be: improving it, or treat it as a reference model for future potential works.

3. Final results reporting

With all results obtained justified and future plans laid out, it is also tine to report them into proper documentations. The final report should consist of all the findings during the development process, as well as anything mentioned previously that is worth an attention. Finally, an overview summarizing the entire project will be included as well to wrap this research project up with a good ending. A final presentation will also be prepared to report the process and findings of this research project that emphasizes on highlights and interesting insight that the panel should know.

4. Project review

Before ending the research project completely, the project should be reviewed by the supervisor and panels to know what would be good to be done again, and what should be avoided. Due to nature of this project being a research project instead of a real-world business use case, there is no review process with clients, managers, team members needed. The project review could also be carried out with friends and potential improvements could be suggested alongside with the things done right or wrong. Overall, the project review is suited for personal benefit instead of entity benefit.

Table 16: Summary of activities and tasks to be done in the Deployment phase

Activities	**Tasks**
Planning deployment	- Results will be written into formal documents - Might be exposed to public through GitHub or journals
Planning monitoring and maintenance	- If results were good, will be used to predict future and identify potential new wave - If results were bad, either improve or use as reference
Final results reporting	- Final report and presentation will be prepared
Project review	- Reviewed by supervisor and panels

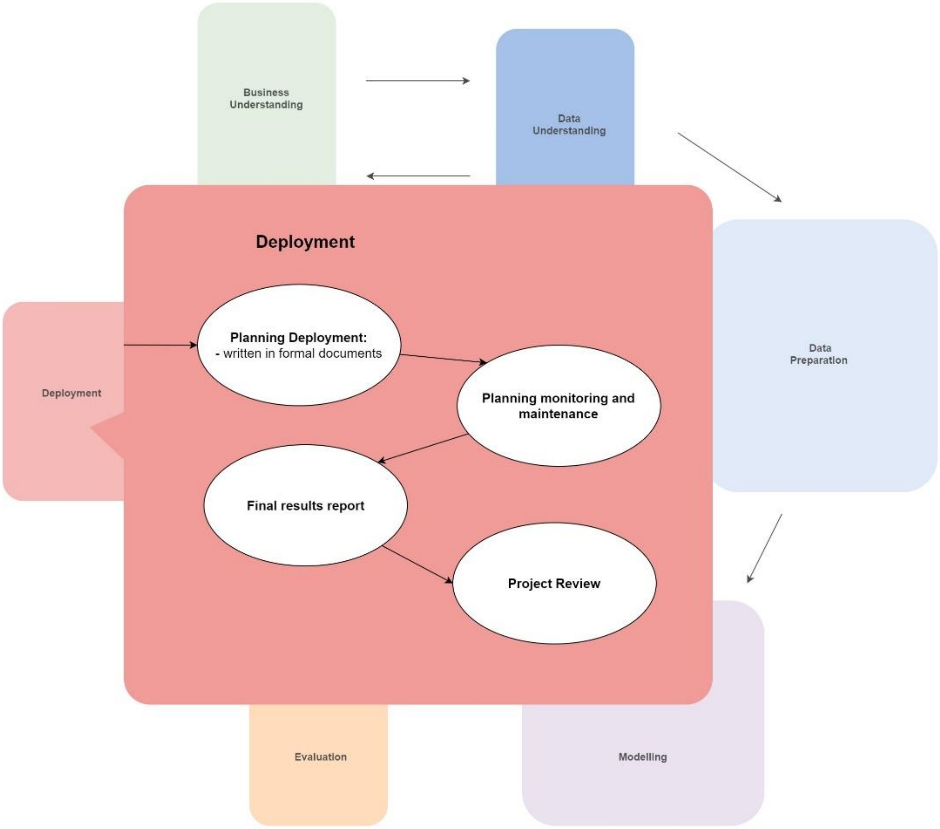


Figure 12: Flow of activities to be done in the Deployment phase

Table 16 shows the summary of activities and tasks to be performed in the deployment phase with the flow visualized in figure 12. It was seen in this phase where the activities were concluded and summarized to an end. The models and results will be deployed into formal documentations such as final report and final presentation while future prospects were also identified based on the final performance of the model. Finally, the deployment phase should end with reviews for personal benefit and marks the end of a formal research project.

3.3 Summary

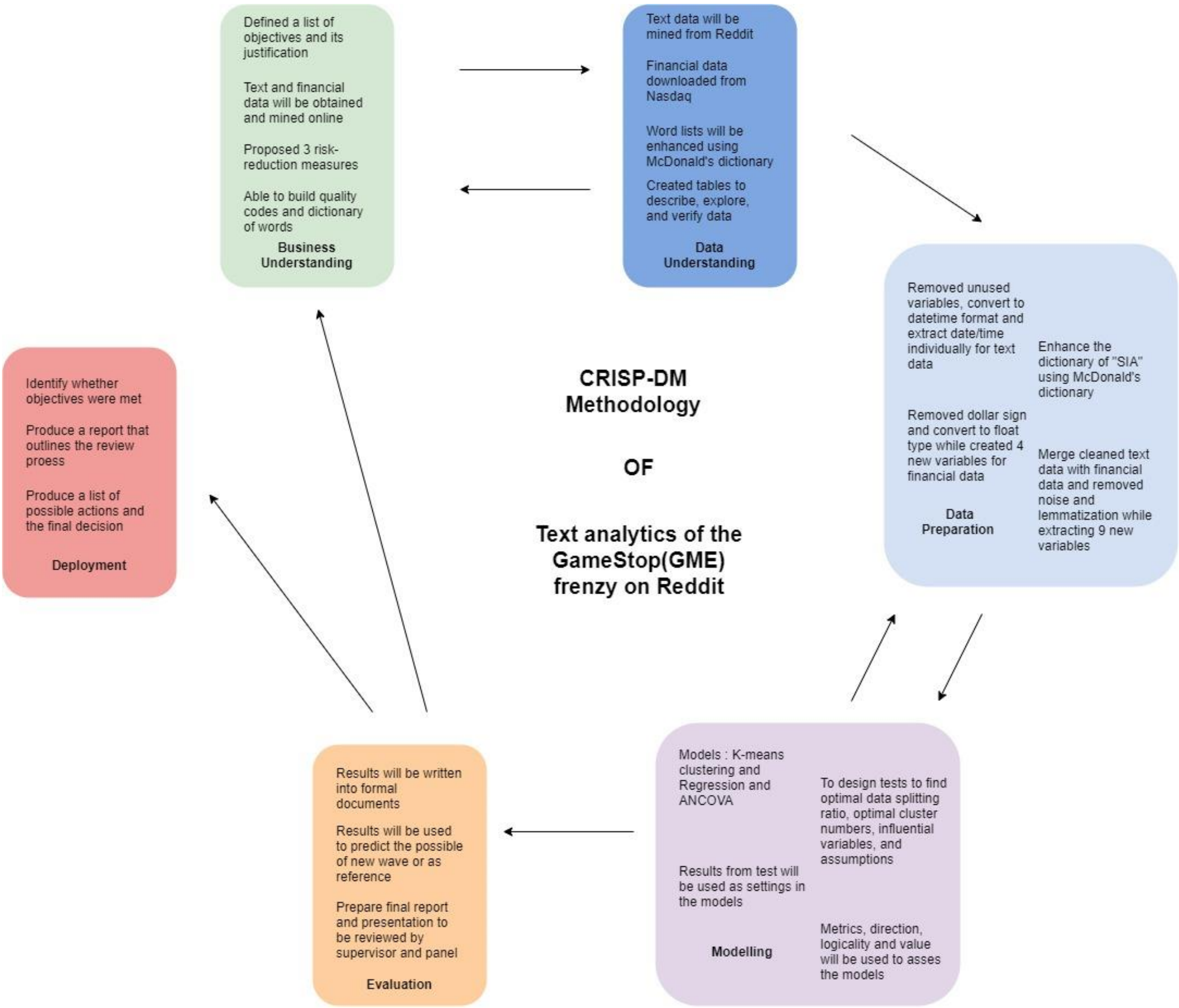

Defined a list of objectives and its justification
Text and financial data will be obtained and mined online
Proposed 3 risk-reduction measures
Able to build quality codes and dictionary of words
Business Understanding
Text data will be mined from Reddit
Financial data downloaded from Nasdaq
Word lists will be enhanced using McDonald's dictionary
Created tables to describe, explore, and verify data
Data Understanding
Removed unused variables, convert to datetime format and extract date/time individually for text data
Enhance the dictionary of "SIA" using McDonald's dictionary
Removed dollar sign and convert to float type while created 4 new variables for financial data
Merge cleaned text data with financial data and removed noise and lemmatization while extracting 9 new variables
Data Preparation
CRISP-DM Methodology
OF
Text analytics of the GameStop(GME) frenzy on Reddit
Identify whether objectives were met
Produce a report that outlines the review proess
Produce a list of possible actions and the final decision
Deployment
Models : K-means clustering and Regression and ANCOVA
To design tests to find optimal data splitting ratio, optimal cluster numbers, influential variables, and assumptions
Results from test will be used as settings in the models
Metrics, direction, logicality and value will be used to asses the models
Modelling
Results will be written into formal documents
Results will be used to predict the possible of new wave or as reference
Prepare final report and presentation to be reviewed by supervisor and panel
Evaluation

Figure 13: Summary of the adoption of CRISP-DM methodology on this project

Figure 13 provides a visualization of the flow and adoption of the CRISP-DM methodology in this research project with lists of tasks to be done concluded during each phase of the methodology. Starting with the business understanding phase, a list of objectives with their justifications were defined to provide project direction and goal. Then, it is identified that the text and financial data will be downloaded and or mined online while proposing 3 risk-reduction measures. With that, data mining goal of being able to produce high quality codes and list of words was identified. With an overview and big picture provided by business understanding, the project moves on to the next phase which is data understanding. It was identified that the text data will be mined from Reddit using their API with Python while the financial data will be downloaded from Nasdaq's website. After that, the word lists will be enhanced using McDonald's dictionary as proposed by many papers and 6 tables were created to describe, explore, and verify data returned from Reddit and Nasdaq as part of the process in this phase. In this phase, as more data were explored and understood, some flaws and potential shortfall might be identified which is why another arrow were pointed back to the previous phase so that the project can still accept adjustments before major work was done. For example, the returned data from Reddit might be insufficient in size and downloading Twitter data will be considered. Then, a return to business understanding phase to is required to provide directions and plans to provide a perhaps new objective or a better goal. However, if everything was laid out clearly and as planned which the revision is not needed. The project could proceed to the next phase of data preparation.

During data preparation, for text data, the unused variables will be removed while converting datetime to a common format for merging later. For the financial data, some data cleaning and manipulation such as removal of dollar ($) sign and the conversion to float type was performed. At the same time, a total of 15 new variables were derived and obtained from the text, date, and prices of stock respectively to enhance the features of the master dataset. Lastly, the word dictionary of "SentimentIntensityAnalyzer"(SIA) was enhanced using McDonald's dictionary to better fit with financial words for higher prediction. That sums up the key tasks in data preparation and the project shall move on to the modelling phase. 3 main models will be carried out for this research project which were K-means clustering, regression, and ANCOVA. Before running the models, a series of tests will be designed and carried out to determine the optimal data splitting ratio, optimal cluster numbers, influential variables and whether the data had met the assumptions for regression and ANCOVA. The results of these tests will be used as the settings in the generation and building of the models. Another arrow was also seen pointed back to the data preparation phase as seen in figure 7 from the modelling phase was identified necessary as the results of the tests might not be ideal especially for the assumptions for regression. With that, the project could revert and perform new round of cleaning and manipulation to ensure that the final data met the assumptions so that an optimal model will be obtained. Finally, once the model was built, different metrics such as accuracy, R-squared, and p-value will be used to evaluate and assess the models with consideration of direction and logicality too. The model and data will constantly be adjusted and edited to obtain a best result from the metrics used which is why revision of the data preparation phase and modelling phase is expected.

After the models were built, the result will be evaluated in the evaluation phase. Firstly, the results will be written and recorded into formal documents for future usage. The end results will be used to predict the possibility of a new wave or as reference only in the future. Then, a final report and presentation will be prepared to be reviewed by the supervisor and panel. Once they are reviewed, there are 2 paths that this project could go on. If the result found were unsatisfied or potential research could be found performed, it could go back to business understanding phase by starting everything all over. With that, a new research project is started, and the CRISP-DM methodology will undergo a new process by itself. If not, this research project could see itself being deployed and shared publicly. A report that outlines the review process will be produced alongside with a list of possible actions to be avoided or removed, with a final decision to be made.

In conclusion, a series of work were planned and scheduled to be done to meet the title and the objectives of this research project. Most of the work was inspired and referenced from other researchers' work which are stated in the literature review to smooth out the process. However, when it comes to the data preparation phase, there is not really 1 source that could be relied on referenced because data cleaning and manipulation is a process that is very case dependent. Thus, more flexibility and revision will be seen during that phase. As for the models, these 3 models were identified initially during project planning. It is also expected to build other models to compare and compare with each other. Thus, many models with different settings and types will be built and the best performing one will be selected to represent the result of this research project. Once the results were obtained, they will be evaluated and revised to obtain a better result. If in the end the results were not ideal enough, this research project will serve as a reference to future works while if an ideal result was obtained, this project will seem to be published and used for the prediction on the possibility of new wave/frenzy. Even though it is not a business use case where deployment was seen crucial, the deployment of this research project will also be seen to provide another evidence and effort to bring people the significance about the power of social media, and people on the stock market.

4. Work Plan

In this section will be discussing about the works to be done in the future. The first deliverable in this section is the activities and work to be done during the commencement in Capstone 2 written in tabular format for easier reading and understanding. Every activity will come with at least one product, risk factors, as well as duration required to complete the respective activity. At the same time, the description of each activity will be included alongside. The second deliverable of this section is the flow chart of work to be performed. It would look similar as to the adoption of CRISP-DM methodology as shown in figure 7 but with more details and tasks to be done. Descriptions will also be provided alongside the flowchart to brief about the flow and plan for easier understanding. Lastly, the final deliverable in this section is the Gantt chart that indicates the planned timeline to carry out the work. The Gantt chart is responsible for the planning of time and schedule for each activity to be performed in Capstone 2 with their expected durations.

4.1 Work Activities

Table 17: Work Activities, risk factors and durations for work plan

Phases	Activities	Product	Risk Factors	Acceptance Criteria	Predecessor/Successor	Duration
Data Understanding	**Collecting initial data:** - Mine data from Reddit - Download data from Nasdaq - Download McDonald's dictionary	A set of raw initial data returned from each activity	Mining too much text data at a time might lead to system crashes	Data that is necessary within the required timeframe is obtained	**Predecessor**: None **Successor**: Data exploration	**1 week** 1 week was allocated because downloading 1-year worth of text data will take some time
	Data exploration: - Explore the data to identify whether is it like planned	An understanding towards the dataset	None	None	**Predecessor**: Collecting initial data **Successor**: Data cleaning / manipulation	**1 day** 1 day is sufficient to explore and understand the data

Data Preparation	**Preparation of text data:** - Remove unused variables according to plan (table 7) - Convert date time to a unified unit - Extract date and time individually as new feature	A new dataset containing necessary data only	Extraction of date and time individually might not be necessary	The dataset is following as planned	**Predecessor**: Data exploration **Successor**: Preparation of master dataset	**7 days** 3 days is sufficient to clean the dataset according to plan
	Preparation of financial data: - Removal of dollar sign - Convert to data type of float for numerical data - Obtain 4 new features: “fluctuation”, “fluctuation_pct”, “returns”, “returns_pct”	A new dataset containing necessary data only	Newly derived variables might not be necessary	The new dataset is successfully obtained as planned	**Predecessor**: Data exploration **Successor**: Preparation of master dataset	**7 days** 3 days is sufficient to clean and derive new attributes according to plan
	Preparation of word list: - Extract positive and negative word list	The word list/dictionary of “SIA” was	The inserted new words might not bring significant	Words were extracted and	**Predecessor**: Data exploration	**7 days**

	- Enhance the dictionary of “SIA” with the extracted word list	enhanced with a new set of financial words	impact to sentiment analytics	loaded inside successfully	**Successor**: Preparation of master dataset	1 day is sufficient to extract and load new words into the python tool
	Preparation of master dataset: - Merge text and financial data - Calculate values for the new 10 features according to plan - Noise removal - Perform lemmatization	A final master dataset with everything combined	The new features might be insignificant and causes computational weight	All data should be filled and meaningful to be analysed	**Predecessor**: Preparation of text data, financial data, and word list **Successor**: Design and run tests	**3 weeks** 3 weeks was scheduled to have sufficient time for cleaning and variable extraction as well as buffer for revision
Modelling	**Design and run tests:** - Test for optimal data splitting ratio - Test for optimal cluster number - Test for significant variables - Test for assumptions	A result that suggests the suitability and settings of data	There will be revision back to the predecessor stages as the tests might not pass well especially for assumptions	All the tests should have a meaningful result and the at most 2 assumptions can allowed to be not met	**Predecessor**: Preparation of master dataset **Successor**: Building models and assessing models	**2 weeks** 3 days will be used to set up the tests and the remaining days till start of week 2 will be used to run the tests. The

						remaining time was allocated as revision back to predecessor for data re-preparation.
	Building models: - Build clustering model - Build predictive model	Models with results will be generated that shows relationship and insights about the data	Many models will be built to find the optimal and satisfying one	Results from test were inserted properly and correctly into the models	**Predecessor**: Design and run tests **Successor**: Assessing models	**1 week** 1 week will be used to build the models using the settings from test while including revision and adjustment of data or settings was required.
	Assessing models: - Assess the models using metrics - Ensure that the result of models fits the direction	Results of the models will be assessed and the best performing will be selected	The result of the models might not be optimal or satisfying even	Model that can sufficiently explain the relationship, trends and	**Predecessor**: Building models **Successor**:	**2 weeks** 1 week will be used to obtain the result of models

	and logicality of the research project		measures were carried out	insights about the data	complete and compile report	and assessed using metrics. The 2nd week is included to revise to predecessors to adjust data and model settings to obtain best performing model if necessary.
Report	Complete and compile report: - Complete documentation of this research project - Perform final compilation and checking with supervisor	Capstone 2 report	Insufficient time	Every detail, steps, and setting applied will be documented	**Predecessor**: Assessing models **Successor**: Report submission, presentation	**3 weeks** 1 week will be used to check the result of the models. The next week will be used to complete the documentation and the last week will be allocated for checking. If

						everything goes right, the submission might be made earlier.
Report Submission + Presentation						

4.2 Flow Chart

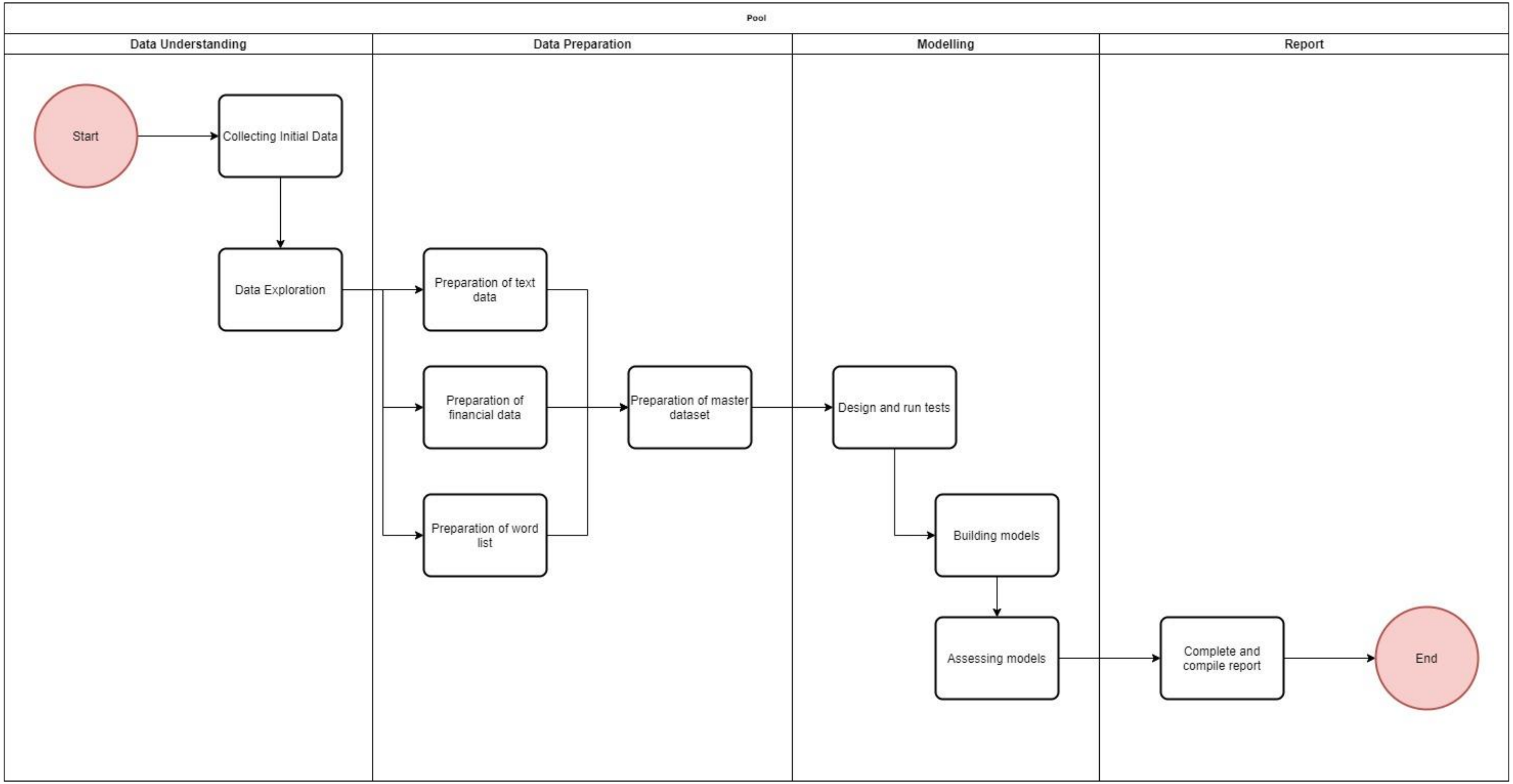
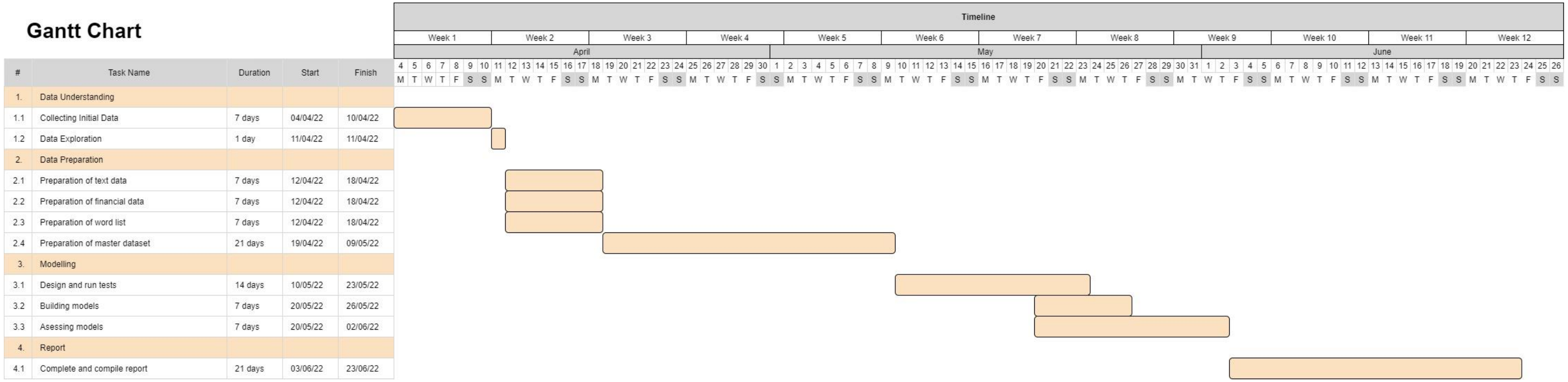


Figure 14: Flow chart of the work plan

4.3 Gantt Chart



Gantt Chart

#	Task Name	Duration	Start	Finish
1.	Data Understanding			
1.1	Collecting Initial Data	7 days	04/04/22	10/04/22
1.2	Data Exploration	1 day	11/04/22	11/04/22
2.	Data Preparation			
2.1	Preparation of text data	7 days	12/04/22	18/04/22
2.2	Preparation of financial data	7 days	12/04/22	18/04/22
2.3	Preparation of word list	7 days	12/04/22	18/04/22
2.4	Preparation of master dataset	21 days	19/04/22	09/05/22
3.	Modelling			
3.1	Design and run tests	14 days	10/05/22	23/05/22
3.2	Building models	7 days	20/05/22	26/05/22
3.3	Asessing models	7 days	20/05/22	02/06/22
4.	Report			
4.1	Complete and compile report	21 days	03/06/22	23/06/22

Figure 15: Gantt Chart of the work plan

5. References

5.1 List of Tables

Table 1: Differences and similarities between short squeezes 7
Table 2: Conclusion of literature review organised following the methodology adopted............. 21
Table 3: Summary of objectives in the view of types of analytics.. 27
Table 4: Summary of activities and tasks to be done in the business understanding phase 32
Table 5: Raw data returned from PushShift API.. 35
Table 6: Raw financial data downloaded from Nasdaq... 37
Table 7: Variable keeping/dropping decision upon exploration of Reddit dataset......................... 38
Table 8: Variable keeping/dropping decision upon exploration of stock dataset........................... 40
Table 9: Data quality of Reddit dataset 41
Table 10: Data quality of stock dataset 41
Table 11: Summary of activities and tasks to be done in the Data Understanding phase 42
Table 12: Summary of activities and tasks to be done in the Data Preparation phase 46
Table 13: List of assumptions for regression and ANCOVA with respective tests 50
Table 14: Summary of activities and task to be done in the Modelling phase 51
Table 15: Summary of activities and tasks to be done in the Evaluation phase 54
Table 16: Summary of activities and tasks to be done in the Deployment phase........................... 56
Table 17: Work Activities, risk factors and durations for work plan 62

5.2 List of Figures

Figure 1: Image representation of the CRISP-DM methodology (Rodrigues, 2020) 9
Figure 2: simple demonstration of the short selling concept 12
Figure 3: simple visualization of the concept of short squeeze 13
Figure 4: Comparison of accuracy between different models for predictive modelling (Mehta, Pandya, & Koetcha, 2021) 19
Figure 5: Comparison between different classification models for predictive models (Joshi, Bharati, & Rao, 2016) 19
Figure 6: Image representation of the CRISP-DM methodology (Rodrigues, 2020) 29
Figure 7: Flow of activities to be done in the Business Understanding phase 32
Figure 8: Flow of activities to be done in the Data Understanding phase 43
Figure 9: Flow of activities to be done in the Data Preparation phase 46
Figure 10: Flow of activities to be done in the Modelling phase 52
Figure 11: Flow of activities to be done in the Evaluation phase 54
Figure 12: Flow of activities to be done in the Deployment phase 56
Figure 13: Summary of the adoption of CRISP-DM methodology on this project 58
Figure 14: Flow chart of the work plan 68
Figure 15: Gantt Chart of the work plan 69

5.3 Bibliography

Anand, A., & Pathak, J. (2021). *WallStreetBets Against Wall Street: The Role of Reddit in the GameStop Short Squeeze.* IIM Bangalore Research Paper No.644.

Awate, A., & Nandwalkar, B. (2019). A Survey on Various methods for Stock Prediction using Big Data Analytics. *Asian Journal of Convergence in Technology*.

Belo, N., Erker, J. J., & Koehler, M. (2021, March 22). Financial Decision Making based on Social Media Sentiment Analysis using Transformers. *Research Paper*.

Bradleya, D., Hanousek, j., Jaame, R., & ZiCheng, X. (2021). *Place your bets? The market consequences of investment research on Reddit's WallStreetBets.* Elsevier BV.

Burnette, R. (2021). *What Were the Factors that led to the GameStop Short Squeeze?* University of Arkansas.

Carvajal, J. A. (2021). *Social media Effects on the market: Reddit Data analysis on Stocks.* Monterrey.

Chung, J. (2021, January 31). *Melvin Capital Lost 53% in January, Hurt by GameStop and Other Bets* . Retrieved from The Wall Street Journal: https://www.wsj.com/articles/melvin-capital-lost-53-in-january-hurt-by-gamestop-and-other-bets-11612103117

Citron Research. (2021, January 27). An update from Citron Research.

Danqi, H., Charles, J., Valerie, Z., & Xiaoyan, Z. (2021). *The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers' Roles in Price Discovery.* SSRN.

Esguerra, H. (2021, January 29). *Meme Stock War Stories: From GameStop Diamond Hands to Robinhood's Reversal*. Retrieved from complex: https://www.complex.com/life/2021/01/gamestop-meme-stock-short-sellers-robinhood

Financial Times. (2021, February 2). *GameStop shares slide 60% as Reddit rally deflates*. Retrieved from Financial Times: https://www.ft.com/content/1be70d9a-91d5-4cbb-a174-37d7b52b6af2

Hailiang, C., De, P., Hu, Y., & Hwang, B.-H. (2011). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*.

Hasso, T., Muller, D., Pelster, M., & Warkulat, S. (2021). Who Participated in the GameStop Frenzy? Evidence from Brokerage Accounts. *Finance Research Letters*.

Hayes, A. (2021, May 31). *How Does the Stock Market Work?* Retrieved from Investopedia: https://www.investopedia.com/articles/investing/082614/how-stock-market-works.asp

Hibbert, A. M., Lawrence, E., & Prakash, A. (2008). *Are Women More Risk-Averse Than Men?* ResearchGate.

JianWei, H. (2015). Online Stock Trading: Do Demographics, Internet Usage, and Attitudes Matter? *International Journal of Business and Social Science, Volume 6, No. 2,* 8 - 15.

Joshi, K., Bharati, & Rao, J. (2016). *STOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS.* Mumbai: arXiv 1607.01958.

Kerckhoven, S. V., & O'Dubhghaill, S. (2021). Gamestop: How online 'degenerates' took on hedge funds. *Exchanges: The Interdisciplinary Research Journal 2021*, 45 - 54.

Lipschultz, B. (2021, March 8). *'Reddit Raider' Favorite GameStop Soars on Latest Cohen Push*. Retrieved from Yahoo Finance: https://finance.yahoo.com/news/gamestop-mania-returns-chewy-cohen-175436958.html?.tsrc=rss

Lipschultz, B. (2021, February 2). *GameStop Rout Erases $27 Billion as Reddit Favorites Tumble*. Retrieved from Bloomberg: https://www.bloomberg.com/news/articles/2021-02-02/gamestop-extends-pullback-with-short-interest-and-volume-sinking

Long, C., Lucey, B., & Yarovaya, L. (2021). "I Just Like the Stock" versus "Fear and Loathing on Main Street" : The Role of Reddit Sentiment in the GameStop Short Squeeze. *SSRN Electronic Journal*.

Lucey, B., & Dowling, M. (2005). The Role of Feelings in Investor Decision-Making. *Journal of Economic Surveys*, 211 - 235.

McDonald, B., & Loughram, T. (2015). The Use of Word Lists in Textual Analysis. *Journal of Behavioral Finance*.

Mehta, P., Pandya, S., & Koetcha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science 7(20):e476*.

Mitchell, C. (2021, May 25). *Short Squeeze*. Retrieved from Investopedia: https://www.investopedia.com/terms/s/shortsqueeze.asp

Monica, P. R. (2021, February 3). *GameStop stock is plummeting but the Reddit rebellion is just beginning*. Retrieved from CNN: https://edition.cnn.com/2021/02/02/investing/gamestop-stocks-investing/index.html

Mox reports. (2018, December 9). *Hedge funds lose $30 billion on VW infinity squeeze*. Retrieved from Mox Reports: https://moxreports.com/vw-infinity-squeeze/

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana. *Applied Computer Systems Volume 25*, 33 - 42.

Palacios, H. G., Toledo, R. A., Hernandez, G., & Navarro, A. A. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Advances in Science Technolog and Engineering Systems Journal*, 598 - 604.

Peiran, J., Andre, V., & Ansgar, W. (2020). *Social Media, News Media, and the Stock Market.* SSRN.

Ponciano, J. (2021, February 11). *Meme Stock Saga Officially Over? GameStop Short Interest Plunged 70% Amid $20 Billion Loss*. Retrieved from forbes: https://www.forbes.com/sites/jonathanponciano/2021/02/10/meme-stock-saga-officially-over-gamestop-short-interest-plunged-70-amid-20-billion-loss/?sh=37d005b0b213

Robinhood. (2021, January 28). *An Update on Market Volatality*. Retrieved from Robinhood: https://blog.robinhood.com/news/2021/1/28/an-update-on-market-volatility

Rodrigues, I. (2020, February 17). *CRISP-DM methodology leader in data mining and big data*. Retrieved from towards data science: https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781

Stewart, E. (2021, January 29). *The GameStop stock frenzy, explained* . Retrieved from Vox: https://www.vox.com/the-goods/22249458/gamestop-stock-wallstreetbets-reddit-citron/

Tolga, B., & Melo, G. d. (2021). Should You Take Investment Advice From WallStreetBets? A Data-Driven Approach. *ArXiv*.

Umar, Z., Yousaf, I., & Zaremba, A. (2021). Comovements between heavily shorted stocks during a market squeeze : Lessons from the Gamestop trading frenzy. *Research in International Business and Finance Volume 58*.

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modeling. *Ecological Modelling*, 312-318.

Wikipedia. (n.d.). *Stock market*. Retrieved from wikipedia.

Winck, B. (2021, January 28). *Robinhood blocks purchases of GameStop, AMC, and others after days of Reddit-fueled rallies* . Retrieved from msn: https://www.msn.com/en-us/money/topstocks/robinhood-clients-say-platform-has-removed-gamestop-and-amc-and-is-only-allowing-holders-to-sell/ar-BB1daWJS