INSY 670 - Social Media Analytics

Final Group Project

MMA3

Leying (Dorothy) Zou ‖ 260950477

Khaled Al-Masaid ‖ 260623070

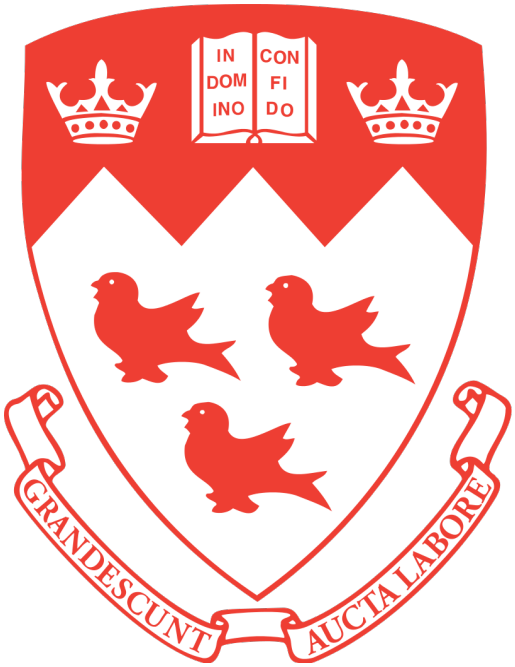
Beiqi Zhou‖ 260742459

Fandi Yi ‖ 260722217

April 13th, 2021

Master of Management in Analytics

McGill University



Presented to:

Prof. Changseung Yoo

Summary

List of Figures

List of Tables

A Context and Problem Statement

Social media's influence in the dynamics of various industries has increased and it is crucial for a company's success to have a positive image on social media. In 2017 United airlines forcibly removed a passenger from their flight and the video circulated on social media resulting in a public backlash against the company which resulted in stock price drop. In February of 2018, celebrity Kylie Jenner tweeted that she no longer uses 'Snapchat' which is another social media platform. As a result of this tweet, Snpachat stock lost $1.3 billion of its value. Most recently, a subreddit called 'WallStreetBets' where participants discuss stock trading gained popularity due its influence on certain stocks that made huge jumps. On January 22, 2021, users of r/wallstreetbets initiated a short squeeze on GameStop, pushing their stock prices up significantly after a comment from Citron Research predicting the value of the stock would decrease. The stock price jumped up significantly by more than 600% at the end of January. As such, the following report examines how social media could change the future of the stock market. How can we use social media data to improve prediction and understanding of the stock market?

The following report explores methods to analyze the effect of social media on stock trends. Social network analysis methods will be used to determine the top influential accounts along with creating network visualization. Sentiment analysis will be performed to determine the correlation between negative sentiment and price drops. Network information along with sentiment analysis will be used to create stock price prediction models.

B Data description

For the purposes of this report, the focus will be on **GameStop** stock that recently sky-rocketed due to social media influence. First, the stock price data for GameStop (GME) were obtained through Yahoo with the following attributes: 'date', 'high', 'low', 'open', 'close', 'volume' and 'adj close'. Then tweets data was scraped from Twitter using the '#gamestop' hashtag through 'snscrape' package on Python from December 27, 2020 till February 26, 2021. The data-set contained 118,000 tweets along with seven attributes including: 'Datetime', 'Tweet Id', 'Text', 'Username','user_mention', 'followersCount', 'listedCount'. This data was then divided into 2 data sets, the complete data with all the extracted tweets was used for sentiment analysis. Duplicated rows have been removed and the tweet texts were cleaned up by removing special characters, and transformed into lowercase. Then, after filtering rows without a NaN in user_mention, this data set was used for network analysis. Similar data pre-processing steps were performed using the "networkx" package, the degree, closeness and betweenness centrality score were computed for each user pair and a correlation map is shown in figure 1. There are are high correlation between listed count and follower count of 0.92, and closeness and degree centrality of 0.97. For the purpose of the stock prediction, listed count and closeness centrality were removed.

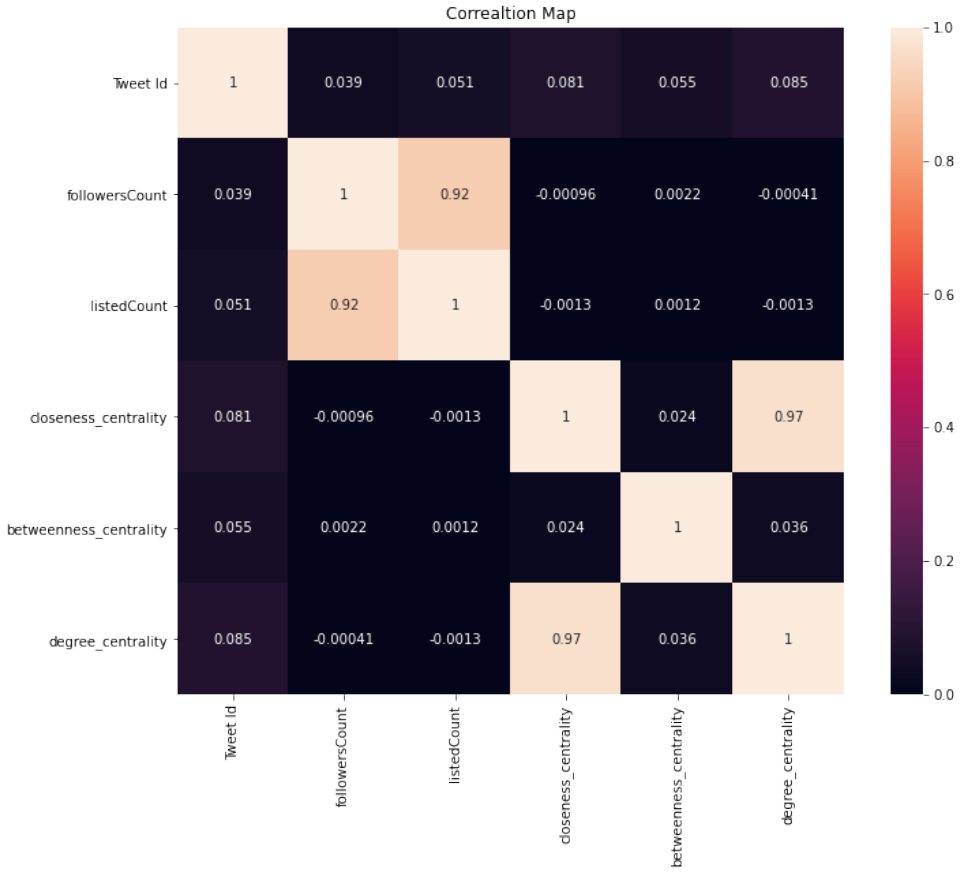


Figure 1: Correlation Map

C Social network analysis

C.1 Methodology

The GameStop's stock big jump happened on January 27^{th}, 2021. Therefore, we divided our network analysis into two parts based on before and after the big jump date. There are four steps to get the final result that include: **data pre-processing, influence score calculation, network plot generation and Investigation of the relationship between Tweet count and Stock price.**

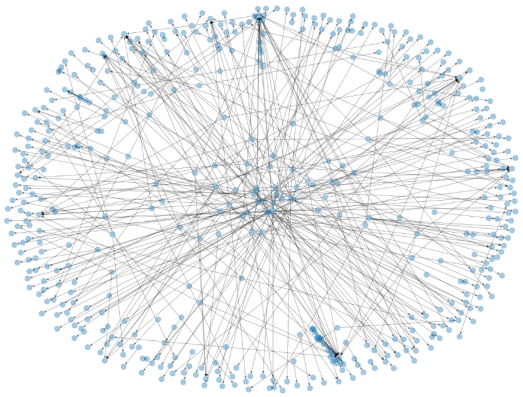
- **Data Processing**: We divided the scraped dataset before and after the big jump event. Then we extended the rows by exploiting the 'user_mention' by setting the one-to-one relationship with 'Username'.

- **Influence Score Calculation**: In order to get the top 100 influencers before and after the big jump event, we first calculated the centralities scores by using `networkx` library, which includes betweenness, degree and closeness. We then applied the weights of the top four important features from part 1 in the previous project that include: **listed_count, network_feature_1, follower_count and network_feature_2**. Besides, we standardized these four features' value before building the equation. We re-weighted the total weight to 1 and build the formula with ***Influencer Score*** = $\sum_{i=1}^{4} Weight_i \times Feature_i$, where i = 1, 2, 3, 4 corresponding to the top 4 features.

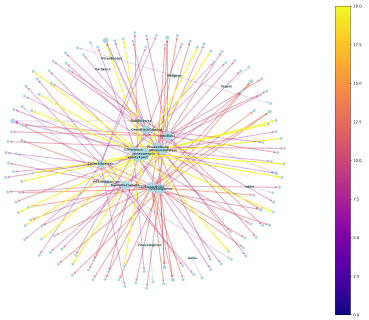
- **Network Plot Generation**: We plotted two different types of the network for before and after stock's big jump by using `networkx` library. The first type of plot is the overall network with all the connections before and after the network. The second type of plot is the top 10 influencers' network, which includes the connections only with the top 10 influencers. In order to create the detailed top 10 influencers' network plots, we set the edge colour and node size based on the network users' degree centrality score.
- **Relationship Between Tweet Count and Stock Price**: The daily amount of tweets were counted across two months and is based on the 'Datetime' column. Then applied the Yahoo stock data to plot the time series line graph for Yahoo stock and Tweet count to investigate the relationship.

C.2 Results

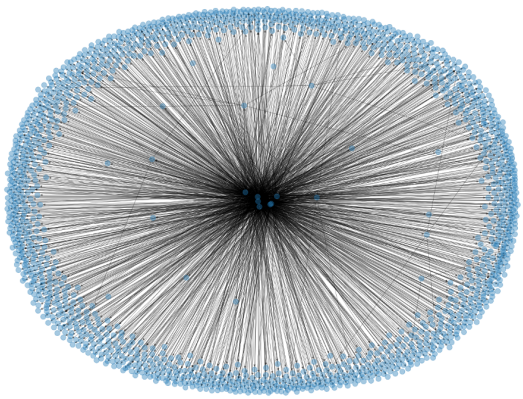
The network plots are shown below (figures 2a and 2c) which display the before and after the big jump of stock price events (January 27^{th} 2021). We got an interesting finding that the network structure has a considerable difference: Before January 27^{th}, the structure is multiple layers which means that most users have equal effect and relatively same amount of connections on '#GameStop' topic. However, after January 27^{th}, the structure becomes more centralized which indicates that a small part of the users made a major part of the effect and a major part of connections.



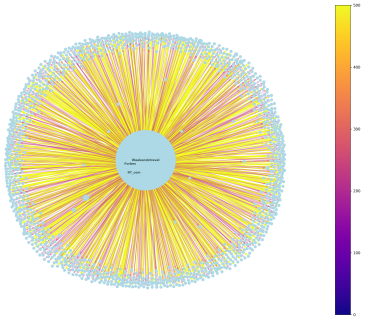
(a) Before Jan 27th



(b) Before Jan 27th top 20



(c) After Jan 27th



(d) After Jan 27th top 5

Figure 2: Network plots

As for the networks shown in figures 2b and 2d, we can have more detailed observations. We plotted the network with the top 10 influencers' connections. The colour of the edge and the node's size indicate the degree centrality score for the pointed nodes. After comparing these two plots, we can see that before Januaray 27^{th}, the nodes' size have small difference and the colour of the edge is varied as well. As for after the big jump event, we can observe that the nodes' size have a considerable difference, and the colour of the edge is relatively consistent.

As the screenshots shown in figures 3a and 3b, in the outcome of those events we checked the user accounts. It can be observed that before January 27^{th}, the users who are a big fan of video games are top influencers, but after January 27^{th} the topic GameStop has drawn attention of more official accounts such as Forbes.

(a) Before Jan 27^{th} **(b)** After Jan 27^{th}

Figure 3: Before and after events

Tweet Count VS GameStop Stock Price: As the figure 4 shown here, the Tweet count has a pretty similar pattern with the stock price, and the peak value happened just one day after the stock's big jump. Therefore, It is a great insight to show that the stock price and Tweet count has a strong correlation in this specific case.

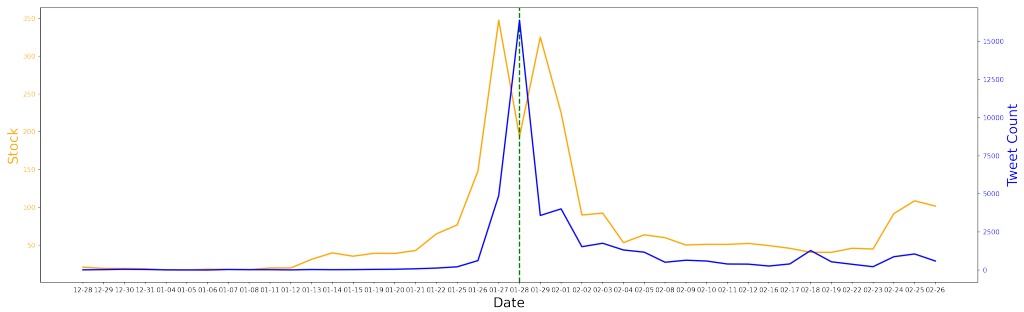


Figure 4: Time-Series Tweet Count and GME Stock Price

D Sentiment analysis

D.1 Methodology

Sentiment analysis played an important role when analyzing the tweet contents. It is efficient with respect to saving time from understanding the sentiment polarity and public opinion manually, so it is a common method in social media analytics and text analytics. In this project, sentiment analysis methodology is used to define the people's attitude to random events such as short squeeze and stock market volatility, and the result can be used in our prediction model to analyze whether the sentiment would impact the stock market or whether the stock market has impacts to sentiment.

The sentiment analysis task of this project included four steps: **data pre-processing, sentiment score calculation, average score calculation, time-series result explanation combined with GME stock price**. The first step is to clean up the tweets because the text content involved some unnecessary information for sentiment analysis. For instance, the sentiment analysis package cannot process emojis yet, so keeping emoji in the text would confuse the model and lead to result inaccuracy. The whole data cleaning function contains the steps of removing the replied messages to someone "RT @", the URL for the tweet, usernames, hashtag symbols, double spacing, punctuations, unnecessary stop words, emojis and numbers. After having cleaned tweets, the next step is calculating the compound sentiment score by using VADER package. It is a library with a trained model to predict the sentiment polarity of text, but it is an unsupervised method to use. The tool will output a score between -1 and 1. Negative scores means negative sentiment prediction and vice versa, and 0 means neutral. Then, the average sentiment score is calculated by grouping up the tweets and their sentiment score by their post date and calculating the average score for each day. The last step is to combine GME stock price data from Yahoo Finances with the average sentiment score per day as a time-series graph. One concern here is that there were no stock price data in weekends because stock market is closed on holidays and weekends. With the dates that we do not have stock data, we eliminated the sentiment result as well to keep the combined data consistent. Thus, we had the sentiment analysis result for each single tweet and the average sentiment scores for each date.

D.2 Results

As shown in Figure 5, it illustrates the time-series changes on average sentiments score and GameStop (GME) stock price from December 27th, 2020 to February 26th, 2021. In general, the average sentiment score for each day showed a **declining trend**. Moreover, the result can be interpreted with stock price in different periods: before the rise, during the climax and cooldown period. At the beginning, the sentiment scores indicated the positive opinions spreading on Twitter about GameStop, and it reached the climax in the test period, which is at the end of 2020. However, the sentiment score started dropping after 2021, and is it is getting close to 0. The stock started rocketing on January 26th, 2021 and the high price lasted a few days. In this period, the sentiment score demonstrated a flat trending with the score close to 0, which means neutral. The stock price had significant reduction started from February and quickly went back to the price which was similar to the price before the rise.

After February, the sentiment score kept showing flat trends with neutral score and the sentiment scored even dropped to negative on two days. It is not easy to analyze the relation between sentiment analysis and price stock with insufficient data; however, figure 5 still reflected people's opinions about this short squeeze. Before the rise, the discussion on Twitter started showing the negative trending. While people from WallStreetBets were overexcited about the event, the average sentiment scores from Twitter were close to 0. There are two possible reasons would lead to this result. The first one is that most of the Twitter users are suspicious to the short squeeze event. Since it is a random event, it is normal for most people to consider it as a high-risk chance. Another possibility is that the numbers of positive tweets and negative tweets are even, so it leads to the average sentiment score close to 0. Furthermore, the low sentiment scores after the stock price decreased also proved that the doubting attitude from public, and it is also possible that Twitter users who are also the investors spread the negative opinions when they lost money from this event. Thus, it can be determined that the sentiment from Twitter is highly correlated with stock price and the sentiment from Twitter was impacted by the stock price as well.

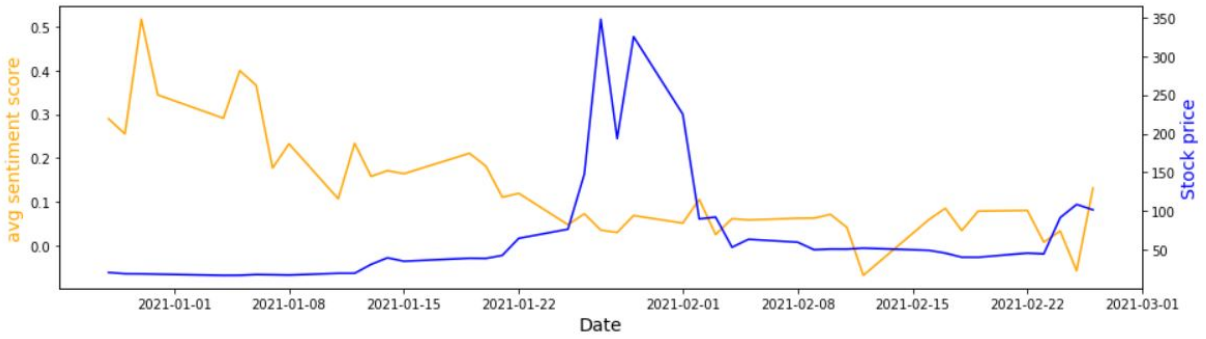


Figure 5: Time-series average sentiment score and GME stock price

E Topic modeling

E.1 Method

Besides social network analysis and sentiment analysis, topic modeling is also useful to extract the trends in data. Topic modeling is one of the most common methodologies for processing a large number of text data, and we implemented its best-known technique called latent Dirchlet allocation (LDA) to discover the topics from the tweet contents (de & Figueiredo, 2021). We used the tools from scikit-learn library, which are feature extraction tool and LDA model to define 5 topics and top 5 words from each topic. The configuration outputs a 5X5 table as the result, which are the topics of the pre-processed tweets regarding to GameStop short squeeze. The method is going to tell us the general topics that Twitter users discuss in the whole test period; however, we are also interested in comparing the topics before and after the stock price increased. Hence, we designed the method that finding the top 1 topic for each date and counting the distributions of different topic before and after January 27th, 2021, and the output will be used to evaluate the impacts of GameStop short squeeze.

E.2 Results

As a result, table 1 displays five topics with top 5 words for each topic from the dataset. The tables involved some words that are less informative such as "di", "und" and "der", but the remaining topics revealed that Twitter users emphasized GME stock, the other discussion network reddit, the trading platform Robinhood and investment trading product hedge funs. These topics are not directly related to the business of GameStop, who is a video game retailer. In other words, the twitter data regarding to GameStop in the test period are mainly about the investment and stock market. Some top words are related to other further stories and news. For instance, the reason of why Robinhood became a popular word is that Robinhood, as one of the biggest trading platform in U.S., restricted the trading of GME and AMC which displeased its users (Winck, 2021).

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
top 1 word	gme	ps	stock	en	hedge
top 2 word	amc	gt	gme	el	die
top 3 word	wallstreetbets	gme	market	reddit	people
top 4 word	gamestonk	link	reddit	street	und
top 5 word	robinhood	di	robinhood	wall	der

Table 1: Five topics with five top words from Topic modeling

Another word called “gamestonk” is from Elon Mask’s tweet, who is extremely influential on Twitter. He posted a tweet on January 26th saying the single word “gamestonk” and the link to WallStreetBets community group in Reddit with 12.2K replies, 46K retweets and 250K likes (Figure 6). Consequently, this word became one of the most popular words in the major topic.

Figure 6: Elon Mask’s tweet about Gamestonk and WallStreetBets

Furthermore, to contrast the topics of tweets about GameStop before and after event, Table 2 and Table 3 are generated by counting the frequency of topic for each date, and they denoted the transfer of topics when people talked about GameStop. Before January 27th (Table 2), people were still talking about games and buying games such as Gran Turismo series game (GT) and gaming equipment such as Play Station (PS) when they mentioned GameStop. However, after January 27th (Table 3), the topics of tweets with GameStop were

only limited to its stock price. Therefore, the topic modeling method signified the topic transferring process on social media, and the top words from topics can interpret the reasons of topic transfer.

Table 2: Top five topics before Jan.27th, 2021

	Topic	Count
top 1 word	gamestop	15
top 2 word	shop	9
top 3 word	gme	3
top 4 word	pt	3
top 5 word	ps	1

Table 3: Top five topics after Jan.27th, 2021

	Topic	Count
top 1 word	gamestop	20
top 2 word	gme	11

F K-means clustering

The K-means algorithm was used to create clusters for tweets based on similarity. The **elbow method** was used to determine the optimal number of clusters and according to figure 7 below, **two** clusters are ideal for this data-set.

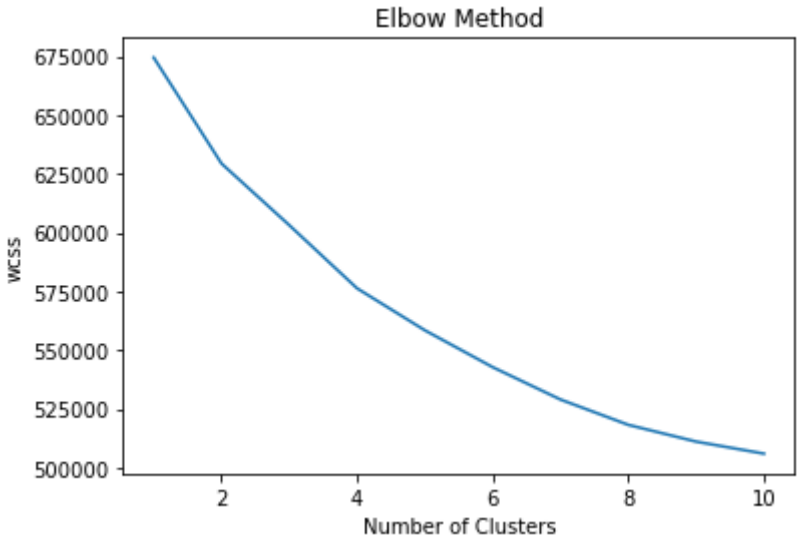


Figure 7: Elbow method

Term frequency-inverse document frequency (TF-IDF) matrix was created and fitted into the K-means model, the results are displayed in figure 8 below. Each cluster can represent users' perception towards the fluctuation of GameStop stock. Since there are only two clusters then the model has classified the tweets into two clusters of opposite sentiments. The results of K-means model creates useful visuals that gives an overview of public's reaction to stock trends and it makes it easier to identify users' perception, in this case it can be observed that clusters are of similar size which indicates there is a balance in users' opinions.

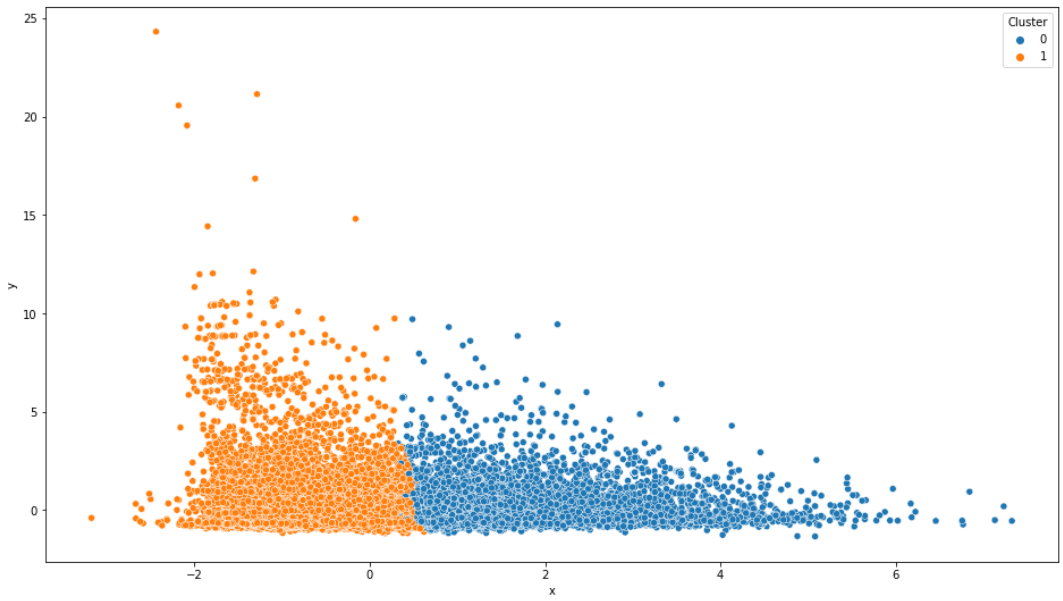


Figure 8: K-mean results for tweet clustering

G Stock prediction

Stock data set downloaded from Yahoo was used to first test the trading strategy based on the predicted result. We used a longer time frame ranging from **2016-02-26** to **2021-02-26**. We have then created new columns based on existing values:

- High minus Low price: High - Low
- Close minus Open price: Close - Open
- Three day moving average: mean 3 days closing price
- Ten day moving average: mean 10 days closing price
- 30 day moving average: mean 30 days closing price
- Standard deviation: for a period of 5 days based on closing price
- Relative Strength Index (RSI): chart the current and historical strength or weakness of a stock or market based on the closing prices of a recent trading period.
- Williams %R: technical analysis oscillator showing the current closing price in relation to the high and low of the past N days.

The target variable (y):

- Price rise: a binary column (0 or 1), 1 if closing price of tomorrow is greater than the today's' closing price

Then, we split the data into test and train set and performed feature scaling. The artificial neural network was built with a sequential classifier. The trading strategy is to take a long position if the predicted value of y is 1 and take a short position if the predicted price is 0. Finally, we computed the returns that the strategy will make at the end of the day based on the actual closing price. The cumulative market returns and cumulative strategy returns are shown in figure 9. Based on the graph, the cumulative market return and strategy return initially followed a similar pattern; however, the pattern stated to get reversed towards the #wallstreetbets event. Only using the historical stock data will lead to a negative cumulative strategy returns, with the highest difference at the end of January. This means that the historical information cannot capture the future movement of the stock price; therefore, we need to take into considerations other factors such as the network and sentiment effect that has initially caused the high rise of the GME stock.

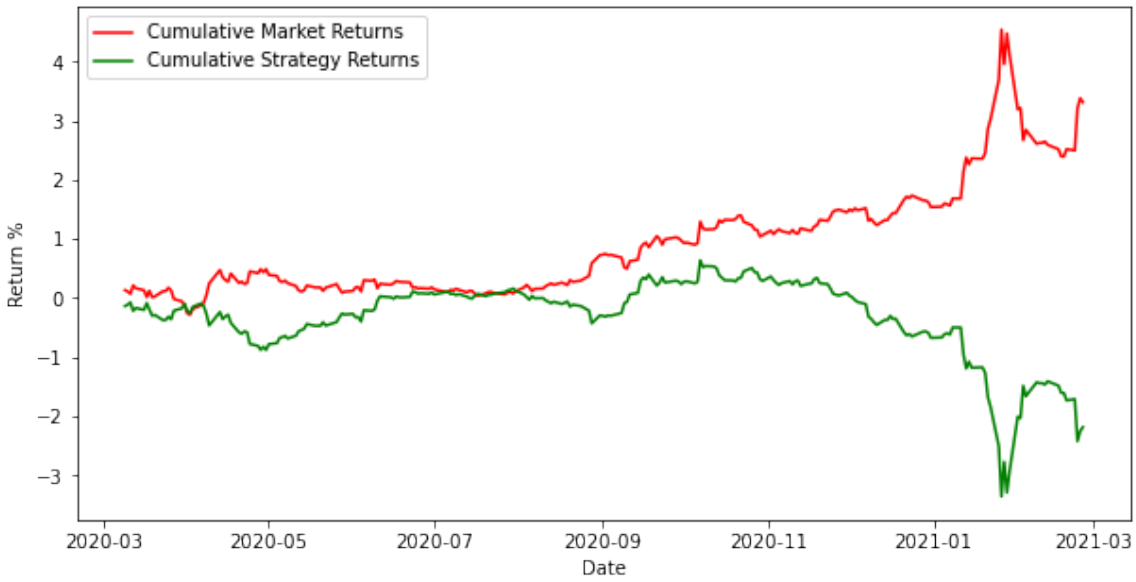


Figure 9: Cumulative stock return

Due to the limited number of total tweets that we can scrap from Twitter and the time limitation, only 3 months of data will be used; however, a longer time frame data will be preferred to have a more accurate result. We have combined the historical GME stock data, the sentiment score together with the network data. The added predictor columns (X) are:

- **Degree centrality, Betweenness centrality, Follower counts, and Sentiment score**

The cumulative return graph using stock, sentiment and network data can be seen in Figure 10a and a same period cumulative return graph using only the stock data can also be seen in Figure 10b for comparison purpose. We can clearly see that the market and strategy return is around 0.8 whereas without the additional data the strategy return would be -0.4 at the end of the period. From this example, we can see that data from social media can be major factor that we need to take into consideration.

(a) Cumulative stock returns(stock, sentiment & network data)

(b) Cumulative stock return using stock data

Figure 10: Comparing stock returns

H Insights & Recommendations

In conclusion, it is evident that social media platforms have an influence on the stock market. When the GameStop stock prices began to soar, users on Twitter had mixed sentiments and there was a balance between opposite sentiments. Network analysis was performed to create visualizations and extract more data regarding the users who tweeted about GameStop in a specific date range. Several conclusion were made from network analysis including the fact that the top influential account after the big jump is a verified account with over 16.7 million followers where as before the jump, the top influential account had only 93 followers. Network information along with sentiment analysis can then be used together to create stock price prediction models with higher accuracy and therefore can lead to higher returns. Predicting the stock prices will allow organizations to make appropriate decisions in anticipation of market fluctuation. Investors should take into account events occurring on social media when making decisions related to the stock market. Businesses need to build a positive image on social media because negative sentiments can lead to stock price drop.

References

[1] de, M. T., & Figueiredo, C. M. (2021). Comparing news articles and tweets about covid-19 in brazil: sentiment analysis and topic modeling approach. Jmir Public Health and Surveillance, 7(2), e24585. doi: https://doi.org/10.2196/24585

[2] Winck, B. (2021, January 28). Robinhood blocks purchases of GameStop, AMC, and others after days of Reddit-fueled rallies. Retrieved from MSN - Business Insider 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

[3] Singh, Devang. (2019, September 05). Neural Network In Python: Introduction, Structure And Trading Strategies: https://blog.quantinsti.com/neural-network-python/?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com