Does Sentiment on WallStreetBets Help to Predict Stock Price?

Social Sentiment Analysis Using Ensemble Approach

Abstract

Sentiment on the WallStreetBets, a Reddit online forum has become a must-track even for mainstream investors due to its role in an unprecedented 2021 rally of GameStop stock. However, limited research exists on evaluating whether WallStreetBets social sentiment meaningfully helps to predict stock prices. In this paper, we predict sentiment over the entire 2021 for a basket of 9 stocks of varying market capitalization using a semi-supervised learning approach. Sentiment features are then used as input to a Long short-term memory model to predict stock prices, along with other financial indicators. Our findings indicate that sentiment does improves the prediction of the some stocks' close price, particularly those of lower market capitalization.

1. Problem Motivation

WallStreetBets (WSB) shot to fame in early 2021 for its role in an eye-popping short squeeze on GameStop (GME) stock that sent the stock from a mere $19 at the start of 2021 up to $483 in end-January 2021. Since then, social sentiment on stock prices has been closely tracked as seen in the proliferation of various sentiment tracking websites like Swaggystocks and Apewisdom. These websites provide high level numbers on overall WSB sentiment and comment volume allowing those who do not have time to pour through the subreddit to get a sense of the topics WSB are actively discussing. However, further study is required to understand if social sentiment is correlated with the stock price and whether sentiment can act as a leading indicator, enhancing related predictive models.

Should a correlation be shown to exist, various financial instruments and trading strategies can be deployed to profitably benefit traders. Conversely, if there is no relation at all, it would imply that the GME rally was a one-off event and sentiment tracking is not at all useful.

1.1 Approach and Methodology

Ensuring that sentiment is correctly captured is critical to enable downstream correlation and prediction analysis. The posts on the WSB forum have a unique syntax and vocabulary which makes interpretation of the call to action and intent of the comments challenging. For this reason, we hypothesize the following:

Hypothesis 1: Automatic sentiment analysis packages will not work well on WSB sentiment.

This is supported by literature which finds that Valence Aware Dictionary and sEntiment Reasoner (VADER) performance comes up to 65% accuracy even after adding WSB-specific terms (AlZaabi, 2021). Another study found that agreement between VADER sentiment and triple-human-annotated sentiment was as low as 10% (Wang & Luo, 2021). Hence, it is expected that using just VADER will be limited in predicting the sentiment of the comments and posts and we should explore alternative approaches.

We thus adopt a semi-supervised learning approach by labelling a random subset of the data to verify the accuracy of sentiment analysis before applying our chosen model to the entire WSB corpus. Besides model choice, Wang & Luo (2021) find that that combining the output of various models (BERT, VADER) can improve performance on the end-task (predicting direction of change in stock price). We hence test all these approaches in isolation before combining features to achieve the best accuracy outcome.

Hypothesis 2: Positive WSB sentiment helps to better predict stock price performance and volatility than overall comment volume.

WSB has an abundance of so-called 'trolls' who post meaningless spam. Hence, we hypothesize that filtering out neutral or noise comments and focusing on positive sentiment signals (in terms of positive comment volume or ratio of positive comments to total comments) may be more useful in prediction compared to total comment volume. To do this, we build a LSTM and establish a baseline accuracy for stock price prediction which includes total comment volume. Sentiment features are then added and incremental accuracy improvement is measured.

Hypothesis 3: The predictive power of sentiment on WSB is stronger for stocks of lower market capitalization.

The narrative of WSB has been one of retail investors taking on traditional institutional investors. In line with this, forums frequently encourage users to band together to influence the price of a single stock. We thus hypothesize that it will be easier for users to influence prices of a stock with lower market capitalization compared to large names,

rendering WSB sentiment useful only if the company is relatively small in terms of market capitalization.

In line with Hypothesis 3, we have chosen 9 tickers with relatively active WSB discussions and categorized them into categories of mega, large, medium and small market capitalizations[1] for the purpose of hypothesis testing.

Table 1. Chosen stock tickers for analysis

TICKER	CAP SIZE	MARKET CAP (FEB 2022)	COMMENTS (2021)
AMZN	Mega	1.56T	11,314
TSLA	Mega	888.82B	51,175
META/FB	Mega	597.60B	5,191
NOK	Large	32.22B	27,000
PLTR	Large	26.31B	57,065
AMC	Mid	9.67B	232,100
BB	Mid	4.06B	191,964
WISH	Small	1.45B	62,550
CLOV	Small	1.23B	127,854

1.2 Adjustments to Initial Proposal

Three key adjustments are made to our initial proposal.

Firstly, instead of multiclass sentiment (positive, negative, neutral) the classification task was changed into a binary one as it was empirically found that negative comments were relatively sparse in the comments (estimated to be <6% of the dataset). WSB by nature is a community where short-sellers or negative sentiment is generally frowned upon unlike other investment communities where balanced advice is preferred. We hence aim to ensure the classifier can distinguish positive sentiment from noise (consisting of both neutral and negative labelled comments).

Secondly, while we had hypothesized that WSB sentiment's predictive power was stronger during meme periods, we have found these periods to be too sparse (<10% of trading days per stock, i.e. <25 days) for meaningful conclusions and remove this hypothesis.

Finally, we have limited our dependent variable to the daily closing stock price of the 9 chosen tickers. Initially, it was proposed to predict standard deviation and direction of change of the stock. However, our chosen approach (LSTM) is inherently not meant to predict these metrics. Furthermore, direction of change is redundant if we can reasonably predict stock price itself. Overall, we find that purely focusing on stock price prediction is sufficient for us to measure usefulness and incrementality of sentiment.

2. Data Collection

2.1 WallStreetBets Dataset

Reddit comment and post data is accessed through Pushshift, an open source API for searching through historical Reddit data. It should be noted that Pushshift is different from the official Reddit API which only provides real-time data. Pushshift data is scraped from Reddit at a point in time and contains noise like deleted comments and posts that are filtered out before analysis.

As we are only interested in sentiment relating to the 9 identified tickers, we first filter for mentions of the tickers in the Reddit post header over the entire 2021. From this we identify 2,682 posts in 2021 where said tickers are quoted and discussed. We then pull all comments associated to these posts, consisting of 1,092,980 comments retrieved. Deleted and duplicated comments are then filtered out, resulting in 766,213 unique comments that are used for sentiment analysis.

2.2 Yahoo Finance Stock Price Dataset

Stock price metrics were derived through the yfinance library which scrapes data from Yahoo Finance through publicly available APIs. Given the previously defined scope, we have 252 trading days' worth of data for the 9 stocks and S&P500.

3. Sentiment Analysis

3.1 Pre-processing

3.1.1 DATA LABELLING

We first sample 5500 comments from 61 posts from the WSB dataset. This is split into batches of 1100 comments per human labeler, and each comment is manually labelled with "positive" (Class 1) or "neutral/ negative" (Class 0) sentiment. To control for subjectivity, each comment is double-labelled, disputes are discussed and resolved with input from the team of 5.

Overall, this provides us with a robust labelled dataset of 5,489 comments (2,658 (48%) positive, 2,831 (52%) neutral/negative) that are used for training and performance evaluation. It should be noted that this sample size is significantly more than comparable studies on WSB (445 (AlZaabi, 2021), 1100 samples (Wang & Luo, 2021), 5000 samples (Alverez et al, 2022)).

3.1.2 DATA CLEANING FOR STATISTICAL BASED MODELS

Prior to passing data into classification models that require text encoding, the following preprocessing steps are taken:

[1] Note that capitalization size is defined according to generally accepted investment community thresholds (i.e. >2BN for small cap, USD 10 BN for mid cap, >USD 200 BN for large cap and mega cap when capitalization exceeds USD 200 BN.

- Lowercase of all text
- Remove any hyperlinks, numbers and punctuation
- Replace emojis with text description from the demoji package
- Remove default stop words and several non-informative Reddit-specific words[2]

3.1.3 Data Cleaning for NLP Based Models

For pre-trained natural language processing models like Valence Aware Dictionary and sEntiment Reasoner (VADER) and Bidirectional Encoder Representations from Transformers (BERT), minimal pre-processing was done as we found the accuracy performance on our labelled dataset to be maximized when pre-processing was minimized. These models are trained on a large corpus, and research (Alzahrani & Johnson, 2021/ Fernandez et al 2022) shows that standard pre-processing applied for statistical based models could be unnecessary and may even negatively impact the performance of the models. Furthermore, these models often have a built-in method for handling emojis. Hence the only pre-processing steps taken are to remove any hyperlinks and numbers.

3.2 Exploratory Data Analysis

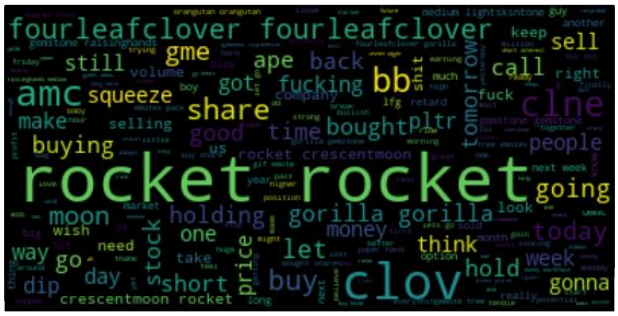

Figure 1 2 *WordCloud of Positive Comments*

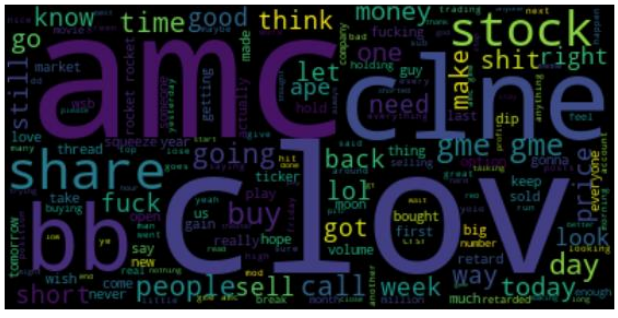

Figure 2 1 *WordCloud of Neutral Comments*

From an inspection of the WordCloud for positive comments (Figure 1) versus neutral comments (Figure 2), visible differences for the top terms emerge. For example, emoji-related terms like the 'rocket', 'gorilla', 'moon' feature highly. Terms relating to potential actions that users may take, for example 'squeeze', 'buying' and 'hold' are also noticeable for positive comments.

For neutral comments (Figure 2) a large proportion of the terms are related to the tickers of the stocks themselves. This may be because many of these comments are speculative, factual, or spam in nature with no clear price support intent in terms of buying or holding. Examples of neutral comments include asking for others' opinion on certain stocks (speculative), users sharing that said stocks hit various price thresholds (factual) or users declaring that they will take certain nonsensical actions if share price for said stock hits a certain threshold (spam).

Overall, this suggests that focusing on a certain number of important terms will be effective in distinguishing positive versus neutral sentiment.

3.3 Statistical based Machine Learning Models

Post pre-processing, the text corpus is fed to TF-IDF vectorizer. The number of terms with highest TF-IDF and n-gram range are hyperparameters which were tuned. It was found that top 250 terms was optimal as balanced accuracy dropped when the number was further reduced. In contrast, balanced accuracy did not drop significantly as the number of terms were decreased from approximately 42,000 terms. On the other hand, changing the n-gram range did not impact balanced accuracy. With inspection of the top terms, it was found that changing the n-gram range from between 1-gram and 2-gram to between 1-gram and 4-gram was not having a significant impact on the top terms selected, since the top terms are more likely to be 1-grams and 2-grams.

The top 250 terms consisting of 1-gram and 2-grams were filtered to form the TF-IDF matrix for model training (*Table 2: Model 1*)

The embedding so obtained contained many features. Dimensionality reduction was performed by both Principal Component Analysis (PCA) and Uniformed Manifold Approximation & Projection (UMAP) to compress the information in the embedding into a smaller, more manageable number of vector components.

Nevertheless, both PCA (*Table 2: Model 2*) and UMAP (*Table 2: Model 3*) performed significantly worse than the baseline TF-IDF model with top 250 terms. For PCA, reducing to 50 dimensions covered only 46% of total model variation, which meant that PCA was not able to effectively reduce the number of dimensions. As seen in

[2] Examples include: 'reddit', 'post', 'comment' which occur with comparable frequency between positive and neutral sentiment.

Appendix 1, UMAP also yielded no interesting segregation between the 2 classes.

3.4 Pre-trained Natural Language Processing Models

3.4.1 VADER

VADER (Hutto & Gilbert, 2014) is rule-based model designed specifically for sentiment analysis and trained on social media corpus like Twitter. One of the key outputs of VADER is a dictionary that matches words to a specific sentiment, ranging from -4 to 4. In prediction, VADER then uses this dictionary to identify the sentiment of the sentence. VADER also takes into account punctuations, letter case, adverb modifiers like "very" and inflection words like "but" that may indicate differing levels of sentiment over and above the sentiment of the word itself. Given that Twitter and Reddit are both social media platforms, it can be expected that the overall grammar and syntax of both platforms may be similar, which rather than standardized English, may include broken grammar, spelling mistakes and wide range of symbols translated as emoticons. Finally, VADER outputs a positive, neutral, negative and compounded final score to indicate the sentiment of the text (*Table 2: Model 4*).

Still, enhancements can be made, when WSB specific terminology and their sentiment is taken into consideration. (*Table 2: Model 5*). Words like "apes", "rocket", "moon" which in real world context usually have neutral sentiment indicate a very strong positive sentiment in a WSB context. Even generally negative words like "retard" have a positive sentiment, given the high degree of self-depreciation noted by Bolyston et al (2021) on the WSB community and is often used as term of familiarity within the community. Therefore, these differences in sentiment were considered and the sentiment scores are updated into the dictionary for sentiment prediction.

Nevertheless, VADER this still has relatively low performance, not exceeding more than 0.62 precision overall. It is hypothesized that there is still a wide range of nuances not captured in the enhanced sentiment dictionary. Hence as shown in *Table 2*, further tuning was done on the cut-off for the point at which the compound final score from VADER was indicating a positive sentiment. A cut-off of 0.6 was chosen it balanced both Precision, F1 score and Balanced Accuracy (*Table 2: Model 6*). In particular F1 was used as a balancing factor as precision continually improves as the cut-off point increases (able to catch all true-positives).

3.4.2 BERT

BERT (Devlin et al, 2018) is a transformer based, deep learning, pre-trained model with almost 110 million parameters, trained mainly on Wikipedia corpus by Google. While the grammar may be more formal, the key strength of BERT lies in the ability to deal with a wide ranging of tasks, with more sophisticated word embeddings (*Table 2: Model 7*).

Nevertheless, a limitation of the dataset is the presence of Out of Vocabulary (OOV) words like stock tickers. While BERT breaks down OOV words into smaller tokens for training, given the limited amount of training data (~4.4k rows), this could impact the results. Hence the removal of stock tickers was tested and Model 8 in *Table 2* shows improved performance.

3.5 Evaluation of Classification models

For evaluating model performance, Precision and Balanced Accuracy were used as the evaluation metrics. Precision was chosen as the cost of false negative is lower than false positives. False positives are expensive because they may directly lead to overly-optimistic assessments of sentiment and ill-advised decisions to buy into the stock. Balanced Accuracy was reported as well for completeness.

Overall, the classification model which performed the best is the Extra Trees Classifier model which utilized a combination of the TF-IDF vectorizer, BERT and VADER outputs. The precision score was 0.90, which was a slight improvement over the next best performance, which was the BERT model with ticker removed (*Table 2: Model 9*).

Table 2. Classification accuracy on test set (1100 comments) of predicted sentiment vs labelled sentiment

	Description	Precision	Balanced accuracy
1	TF-IDF (LGBM)	0.71	0.69
2	TF-IDF (LGBM) + PCA	0.64	0.66
3	TF-IDF (KNN) + UMAP	0.59	0.59
4	VADER BASELINE	0.48	0.53
5	VADER (+WSB Vocab)	0.54	0.62
6	VADER SENTIMENT + PRECISION TUNED	0.62	0.64
7	BERT BASELINE	0.83	0.84
8	BERT (Ticker Removed)	0.89	0.93
9	TF-IDF + BERT + VADER OUTPUT (EXTRA TREES)	0.90	0.90

These results match up well to benchmarks in existing literature that have used similar human labelled data. For VADAR performance, maximum plain accuracy is around 0.65 (AlZaabi, 2021) versus the 0.64 balanced accuracy we have achieved. For BERT performance, precision for the majority class is found to be 0.89 (Alvarex et al, 2022), remarkably similar to the 0.89 precision for the positive class which we have achieved here. This gives us confidence that each approach we have tried is generally performing to the maximum it can be expected to achieve.

Hence, we choose to proceed with *Model 9* based on precision as the cost of false negative is much lower than false positives as discussed above. The best performing model was then utilized to label the rest of the dataset which had not been labelled by a human labeller to

generate the sentiment related features discussed later in the report.

4. Stock Price Prediction

4.1 Regression Data Setup

Features used for regression are set up on a daily basis for the entire 2021 (252 trading days). The dependent variable is the Daily Close Price (*Close*) of the stock. Features used for prediction are discussed below.

4.1.1 Sentiment Features

We engineer the daily count of positive and neutral comments as well as the percentage of positive comments in the following way. With sentiment predicted for all comments in our dataset, we assign each comment to a stock ticker using the following attribution logic:

- Count mentions of our 9 specific tickers within a certain comment (e.g. 'bb' or 'blackberry'). The comment would be assigned to the ticker with the highest count of mentions in the comment.
- Count mentions of our 9 specific tickers within the post the comment is reacting to.
- In the situation the comment does not mention a stock (count for all tickers is 0 in the first step), it will be assigned to most mentioned ticker in the post the comment is reacting to.

This approach to search for short-form tickers to attribute comments is similar to that taken by Buz & de Melo (2021) and found to be more comprehensive than other methods of filtering such as only looking for abbreviations that only contain a preceding '$' sign.

Following this, positive and neutral comments are aggregated by day. This gives us 3 sentiment features: *positive comment count*, *neutral comment count* and *percentage positive comments* (computed as positive comments divided by total comments). Moving averages and moving sums of positive, neutral and total comments are created from 2 up to 7 days to provide a smoothened out time series. This gives us a total of 42 sentiment features per stock (3 original features × 2 moving computations × 7 time ranges). While these are highly correlated, our forecasting approach (LSTM) can handle multicollinearity.

4.1.2 Baseline Features

Baseline features were created in the following categories:

- Basic Stock Price Information (5 features): *open*, *high*, *low* prices, *volume* traded for that day and intra-day standard deviation *(SD)*
- Change Information (14 features): In line with the literature (Asness et al 2014) two momentum indicators were generated for the stock price dataset. Percentage change (*%Change*) and direction of change (*Dir* – binary variable representing whether the stock increased (1) or decreased (0)) is engineered. Both are calculated over 7 preceding days.
- Comment volume (15 features): As we want to measure the incremental impact of adding sentiment, *total comment count* and respective moving sums and averages are added as this information is available irrespective of sentiment analysis. An additional binary variable (is_meme) is also engineered to capture accelerating interest in the stock. This is calculated as comment volume exceeding more than two times the 30-day moving average, and in the 90th percentile of daily comments so far.
- Others (1 feature): *Day of week* is also added as a categorical variable.

Note that while there are only 252 trading days, social features such as sentiment and comment volume (discussed below) is available 365 days a year. To address this, when applying time lags (section 4.3.2), social data is shifted according to calendar date while financial data is shifted according to trading day. For example, if time shift is 1 day, Sunday's sentiment will be used to predict Monday's close price. In contrast, as there is no financial data for Sunday, Friday's financial data will be used to predict Monday's close price.

4.2 Exploratory Data Analysis

Visual inspection (*Figure 5*) shows that for certain stocks (e.g. PLTR), there is some evidence that spikes in positive comments are closely followed by increases in price. However, the relationship is not straightforward across all stocks and will be demonstrated more concretely through improvement in prediction performance.

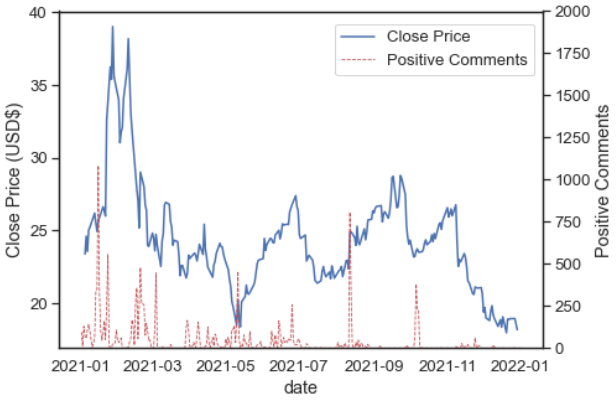

Figure 5 *PLTR close price vs positive comments over 2021*

4.3 Regression Approach

4.3.1 Model Choice: Long-Short Term Memory

The Long-Short Term Memory (LSTM) recurrent neural network model is commonly used to predict future stock market values in academia (Adil Moghar, 2020) as

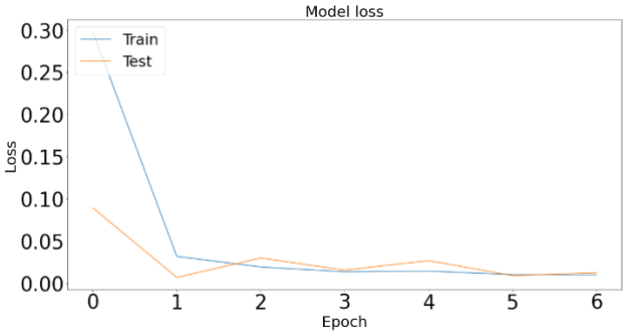

Figure 6 *Training and validation loss values for AMC at 1 time shift (data not normalized and excluding sentiments)*

financial forecasting is regarded as a time-series. Other advantages of the LSTM model include insensitivity to gap lengths as the gates can retain relevant information over a longer period of time, insensitivity to multicollinearity, and partial effectiveness against the vanishing gradient problem with direct access to the forget gate activations.

Since we are less interested in the prediction actual stock prices, and more focused on testing the effect of sentiment on stock price prediction performance, particularly if the stock have lower market capitalization, we have run separate prediction models, one with sentiment data, and the other without, *ceteris paribus*. Any difference in prediction performance can then be attributed to the presence (or absence) of sentiments.

Like most neural networks, LSTM works best with data between 0-1, hence min-max scaling was applied. It is believed normalizing will improve the efficiency and accuracy of training since LSTMs use small weight initialization, and unscaled data may result in the model learning large weight values, causing instability and higher generalization error. Hence, we compared models with only min-max scaling applied, as well as models with an additional standard scaler applied to evaluate if normalization improves predictive performance. Subsequently, we observed that normalization did not improve performance across the board, contrary to common practice (Stottner, 2019). It is likely that normalizing after applying min-max diminished its intended effectiveness since all features are kept within a smaller and only positive range, making it harder for the nodes in the layers to learn. Furthermore, for the retail investor to be able to make beneficial decisions, there needs to be larger variances detected in the trend to profit from, which the normalization would inadvertently have smoothened. Hence, our final model did not include any normalization of data.

Our network consists of four layers, the first is an LSTM layer which takes each mini-batch as input. Each mini-batch has 7 steps and 100 variables, creating 700 neurons. The second LSTM layer's input is the sequence from the previous layer and returns 5 values, the third layer is dense with 5 neurons and the final dense layer outputs the predicted dependent variable. Different activation functions were attempted: tanh, sigmoid, ReLu and softmax, with the lowest validation loss observed with tanh and sigmoid. Since the sigmoid function is slightly more prone to the vanishing gradient problem, we went with tanh, which is less prone as values are centered at zero.

In modelling, we have employed early stopping to prevent overfitting on the validation set to allow for optimal generalization performance. Patience is set as 5, since it is observed most models stop training when epochs < 20. A lower patience would mean earlier stopping, resulting in an insufficiently trained model, a higher patience may result in overfitting, as the training and validation loss plots occasionally show spikes from one epoch to next as seen in *Figure 6*.

4.3.2 Prediction Horizon & Lag Periods

As we are investigating the effect of sentiment of stock performance, we expect that sentiment changes quickly in a short period of time. Hence, each mini-batch will consist of 7 sequences (7 days or 1 week), where the sequence is the 20% test set. A rolling window approach is used where if the sequence length is 25, then the input of days 1-25 (inclusive) will be used to generate a prediction for day 26. *Figure 7* shows the data structure just described.

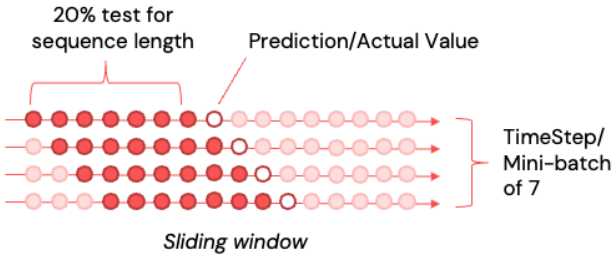

Figure 7 *3-dimensional data structure*

Furthermore, to test robustness, we train the LSTM at lags of 1, 4 and 7 days to observe if sentiment is accretive to prediction outcomes across various advance time horizons.

4.4 Results Interpretation

4.4.1 Evaluation Metrics

Mean Absolute Percentage Error (MAPE) and Median Absolute Percentage Error (MDAPE) are suitable metrics for evaluation of the regression results. Being expressed as a percentage, these two metrics can be used for comparison of the regression results between different stocks, as is necessary for testing Hypothesis 3.

As MAPE relies on mean instead of median in the calculation, MAPE would be more affected by outliers

compared to MDAPE. Considering that the prediction is potentially used to make stock trading decisions, we consider MAPE to be the more appropriate metric because prediction 'outliers' can potentially cause huge losses and should be considered as part of the evaluation metric.

4.4.2 General Interpretation

Table 5 below summarizes the full results for 3 models with different lag periods tested (1, 4 and 7 days) by stock.

Firstly, this shows that the LSTM MAPE is generally low and robust for mega-cap, large-cap and mid-cap stocks, all of which have MAPE below 16% across all time periods. This may also be because small-cap stocks tend to be exceedingly volatile.

Secondly, it shows that LSTM MAPE generally increases as prediction horizon increases, with the 1-Day time shift generally having the lowest MAPE. This is to be expected as recent price should generally have the least fluctuation from current price to be predicted.

Table 5. Full MAPE Results for different lag periods

Ticker	Scenario	1 Day	4 Days	7 Days
AMZN	Baseline	2.2	2.3	2.4
	+Sentiment	2.9	3.1	3.7
	Improvement	*0.7*	*0.8*	*1.3*
TSLA	Baseline	9.9	10.8	9.4
	+Sentiment	7.4	8.6	8.3
	Improvement	*-2.6*	*-2.2*	*-1.0*
FB	Baseline	3.7	3.3	4.4
	+Sentiment	2.9	5.2	6.9
	Improvement	*-0.8*	*1.9*	*2.6*
NOK	Baseline	3.2	2.2	3.0
	+Sentiment	4.0	8.3	8.6
	Improvement	*0.8*	*6.1*	*5.5*
PLTR	Baseline	4.4	7.6	10.1
	+ Sentiment	4.3	7.2	6.7
	Improvement	*-0.1*	*-0.4*	*-3.4*
AMC	Baseline	9.3	10.6	11.9
	+ Sentiment	6.9	12.4	15.6
	Improvement	*-2.4*	*1.8*	*3.8*
BB	Baseline	4.9	5.5	5.5
	+ Sentiment	5.3	5.1	5.7
	Improvement	*0.5*	*-0.4*	*0.2*
WISH	Baseline	29.0	34.6	31.9
	+ Sentiment	15.0	15.6	15.3
	Improvement	*-14.0*	*-19.0*	*-16.6*
CLOV	Baseline	106.7	444.1	418.7
	+ Sentiment	19.1	19.2	27.0
	Improvement	*-87.6*	*-424.9*	*-391.7*

4.4.3 Impact of Sentiment

Table 6 further summarizes the results of the regression performance in *Table 5*. It can be observed that MAPE for 7 out of 9 stocks, shown an improvement in MAPE after addition of sentiment features. Furthermore, sentiment features definitively improved the prediction results for 4 stocks, namely TSLA, PLTR, WISH and CLOV as results improved over all lag periods.

However, sentiment is not always informative, as 2 out of 7 stocks (AMZN and NOK) showed no improvement on MAPE across all time periods.

Table 6. Summary of models with best MAPE improvement

Ticker	Best Lag	Baseline	With Sentiment	Change in MAPE
AMZN	1 day	2.2%	2.9%	0.7%
TSLA	1 day	9.9%	7.4%	-2.6%
FB	1 day	3.7%	2.9%	-0.8%
NOK	1 day	3.2%	4.0%	0.8%
PLTR	7 days	10.1%	6.7%	-3.4%
AMC	1 day	9.3%	6.9%	-2.4%
BB	4 days	5.5%	5.1%	-0.4%
WISH	4 days	34.6%	15.6%	-19.0%
CLOV	4 days	444.1%	19.2%	-424.9%

As for the impact of market capitalization, in line with Hypothesis 3, the two stocks categorized as small capitalization stocks (WISH and CLOV) showed the greatest improvement overall, with MAPE reducing by -19% and -425% respectively. This improvement in the forecasting of WISH's stock price is displayed clearly in Appendix 2. Even so, the relationship with capitalization is not straightforward for large and mega cap stocks. We would generally expect to find low to no impact for these categories, but sentiment has still proven useful especially for TSLA and PLTR.

Finally, we find that lag period is important in that MAPE improvement for sentiment does not always appear strongly in all time periods. In particular, for 3 stocks (FB, AMC and BB), the best improvement appeared at a lag period of 4 days. For stocks like PLTR, the best lag was as late as 7 days. If lag period other than the one stated in *Table 6* was chosen, sentiment may have a negative impact on the prediction performance. Overall, this suggests that sentiment information may take some time to affect stock price and that looking at longer time lags (4 to 7 days) may be beneficial.

5. Conclusion

5.1 Summary of Results by Hypothesis

Through the study, we have made significant contribution towards the 3 key hypotheses in Section 1.1.

Hypothesis 1: **Efficacy of sentiment analysis models on WSB data** - Our results for using a pure sentiment model (VADER) relatively poorly even with fine tuning to include WSB specific jargon. This is similar and to be expected from existing literature. However, when combining TF-IDF features, VADER output and BERT output and applying a final classifier, we can still improve the accuracy of the prediction of double-blind labelled sentiments significantly to achieve an excellent precision result of 0.9.

Hypothesis 2: **Impact of using sentiment to predict stock prices** – Our LSTM model shows that sentiment in general can help in the prediction of short-term stock prices versus a baseline model that considers mainly financial indicators, using information available for the last 7 days. For TSLA, PLTR, WISH and CLOV in particular, WSB sentiment provided useful information that can be used to predict the stock's close price in the following period. This could be because these stocks' investors were influenced by WSB comments or the investors were making positive comments on WSB prior to their purchase decisions. Alternatively, it could also be due to WSB comments being representative of the investors' thoughts.

For the other stocks, comments on WSB were not definitively useful. This may be due to commenters not taking investment actions based on their comments or majority of the price action is not influenced by retail investors anyway.

Hypothesis 3: **Impact of using sentiment based on market capitalization** – Sentiment is particularly useful when applied to smaller market capitalization stocks like WISH and CLOV. However, results here are only for two small cap stock and future testing should focus on verifying this conclusion over a wider pool of small cap stocks.

Overall, even though results for Hypothesis 2 and 3 are mixed, this study still suggest that investors should not ignore the impact of sentiment and that tracking can be useful in some cases.

5.2 Application to real-world problems

Sentiment information is becoming increasingly ubiquitous and is a rich and (often) free data source. Hence, it is important to understand the extent to which such information can be leveraged to better predict future outcomes. While the causal mechanisms are not always clear, this does not detract from the potential usefulness of sentiment as an input to various types of price and demand forecasting models.

Our project can be applied to any context where the opinions of a large group of participants is an important input. For example, besides stocks, it can apply to speculative assets like cryptocurrencies, or even stable assets like property prices as price fluctuations in these assets are largely influenced by the broader market rather than institutions. While the financially literate advise against emotional investing, the converse is true where it is possible to profit off the poor investment decisions driven by emotions of the other parties in the market. Top-performing funds outperform because they can exploit sentiment changes and profit from arbitrage (Yong Chen, 2021). Hence it is important to understand common sentiment not necessarily to join the rally, but rather to stay ahead of it.

One of the difficulties of extracting sentiment is not just the sheer volume of a large corpus, but also when said corpus is highly specialized, as is the case for WSB. The approach that had been adopted in our study for sentiment analysis could be adopted where sentiments are exchanged on a platform which has its unique syntax (e,g EDMW HardwareZone, 4Chan, Hive). The combination of pre-trained models such as VADER and BERT with statistical methods such as TF-IDF vectorizer resulted in a good performance which allowed the other prediction tasks to be executed.

Sentiment analysis on a time series can also be used in non-financial applications such as political elections or in forecasting demand for a product to then conduct the necessary supply planning. For instance, understanding sentiment towards Covid-19 as the pandemic evolved to a pandemic over time can help healthcare systems and pharmaceuticals forecast the demand for vaccines and make the appropriate logistical arrangements to ensure supplies meet demand. In the case of other consumer goods, this forecast can also be used to make pricing decisions and drive margin growth.

5.3 Limitations and Improvements

There are 3 key limitations and potential for future improvements to be highlighted.

Although the amount of manually labelled data is higher compared to studies sampled in academia, it is still a small portion over the entire size of data. Given the wide range of emotions conveyed, the distribution in the sample data may still be different from the entire population. In particular, only ~35% of the larger data set was noted to have a positive sentiment, compared to 48% in the labelled data set. One way to address this could be to use the labelled data, use clustering methodologies to identify neighbours to expand the labelled data set, before training the sentiment prediction models on this new labelled

dataset. A more reliable method would be to simply have more labelled data, which would take more time.

As noted in Section 2.1, only 766K comments were used. About 26% (282K comments) were deleted before they could be captured by PushShift API for archiving, returning only a '[removed]' description upon retrieval. These removed comments may have been removed because they were more controversial or violated community guidelines. However, they could also have contributed to the sentiment of the community at the point in time and therefore limit the study from capturing the true sentiment at the point in time. While this cannot be addressed retrospectively, any further study to examine the correlation between WSB community sentiment should actively scrape live data from the official Reddit API over a period time to ensure completeness of information.

Finally, it should be noted that 2021 was a bull market (Rosenbaum, 2021/ Forsyth, 2021) with exceptional returns. This could have also better supported any positive sentiment and speculative trading. The correlations observed in the study may be weaker in bear markets. Studying the trend over a longer time period is needed to established a stronger relationship between WSB sentiment and stock prices and may well help to establish if there is a meme effect for stock prices in specific periods where there is significant hype over the stock.

6. Supporting Code

Supporting code for this project may be found at https://github.com/rachelsng/WallStreetBets-Sentiment .

7. Appendix

Appendix 1 - UMAP Visualisation

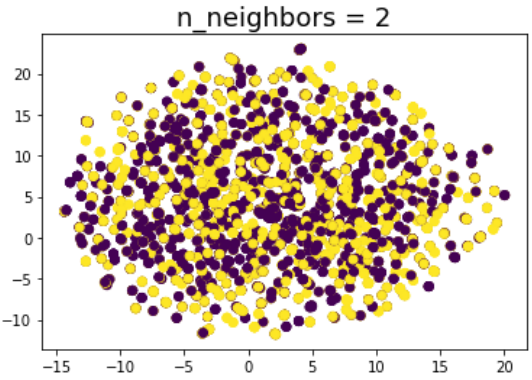

Figure 8 *UMAP Mapping; N Neighbours = 2*

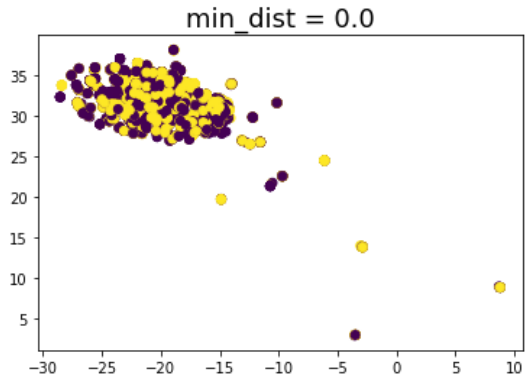

Figure 9 *UMAP Mapping; Min Distance = 0*

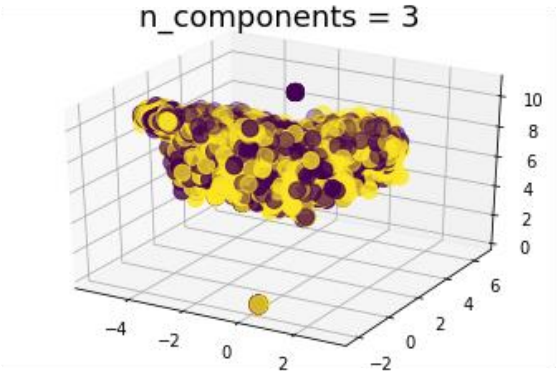

Figure 10 *UMAP Mapping; Components = 3*

Appendix 2 – Results for WISH prediction with and without sentiment features

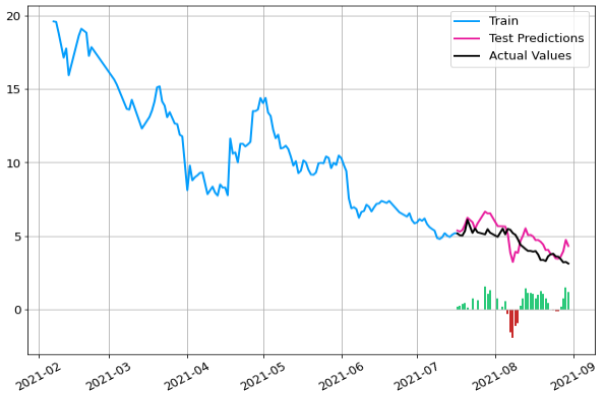

Figure 11 *WISH close price prediction WITH sentiment features*

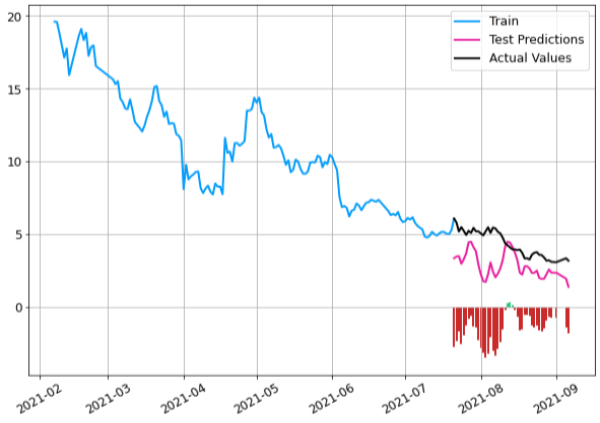

Figure 12 *WISH close price prediction WITHOUT sentiment features*

References

Adil Moghar, M. H. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. Procedia Computer Science, Volume 170, 1168-1173. Retrieved from https://doi.org/10.1016/j.procs.2020.03.049

Alvarez, R., Bhatt, P., Zhao, X., & Rios, A. (2022). Turning Stocks into Memes: A Dataset for Understanding How Social Communities Can Drive Wall Street. *arXiv preprint arXiv:2203.08694*.

AlZaabi, S. A. (2021). Correlating Sentiment in Reddit's Wallstreetbets with the Stock Market Using Machine Learning Techniques.

Alzahrani, E., & Jololian, L. (2021). How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors. arXiv preprint arXiv:2109.13890.

Boylston, C., Palacios, B., Tassev, P., & Bruckman, A. (2021). WallStreetBets: Positions or Ban. arXiv preprint arXiv:2101.12110.

Buz, T., & de Melo, G. (2021). Should You Take Investment Advice From WallStreetBets? A Data-Driven Approach. *arXiv preprint arXiv:2105.02728.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fernández-Martínez, F., Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., & Montero, J. M. (2022). Fine-Tuning BERT Models for Intent Recognition Using a Frequency Cut-Off Strategy for Domain-Specific Vocabulary Extension. Applied Sciences, 12(3), 1610.

Forsyth, R. W. (2021, December 27). What will it take to kill this bull market? we'll find out soon. What Will It Take to Kill This Bull Market? We'll Find Out Next Year. | Barron's. Retrieved April 21, 2022, from https://www.barrons.com/articles/bull-stock-market-rally-51640273934

GitHub - pushshift/api: Pushshift API. GitHub. Retrieved April 19, 2022, from https://github.com/pushshift/api

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

Long, C., Lucey, B. M., & 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, 31, 1-37.

Rosenbaum, E. (2021, December 31). The Bull Market's biggest hopes for 2022 are in the portfolios of wealthy young investors. CNBC. Retrieved April 21, 2022, from https://www.cnbc.com/2021/12/30/bull-markets-biggest-hopes-for-2022-rest-with-millennial-millionaires.html

Stottner, T. (2019, May 16). Why data should be normalized before training a neural network. Medium. Retrieved April 23, 2022, from https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d

Wang, C., & Luo, B. (2021). Predicting $ GME Stock Price Movement Using Sentiment from Reddit r/wallstreetbets. In Proceedings of the Third Workshop on *Financial Technology and Natural Language Processing* (pp. 22-30).

Yong Chen, B. H. (2021, August). Sentiment Trading and Hedge Fund Returns. The Journal of Finance, 76(4), 2001-2033. doi:https://doi.org/10.1111/jofi.13025