Results

1. Data Exploration

1.1 Engagement by day of the week

Figure 1 shows an extraordinary leap in engagement on Fridays. This on the outset would appear odd – the stock has experienced white knuckle volatility on all weekdays. One would expect the levels of engagement to reflect that with a more even distribution across weekdays. One explanation could be that there is a greater level of positivity (and intoxication!) on Friday that encourages more engagement. Another interesting theory is related to options (i.e., calls and puts) expiring on the third Friday of every month. This could be a result of using bot farms to influence investors and protect assets, although this assertion is little more than a conspiracy theory and should be verified in further work.

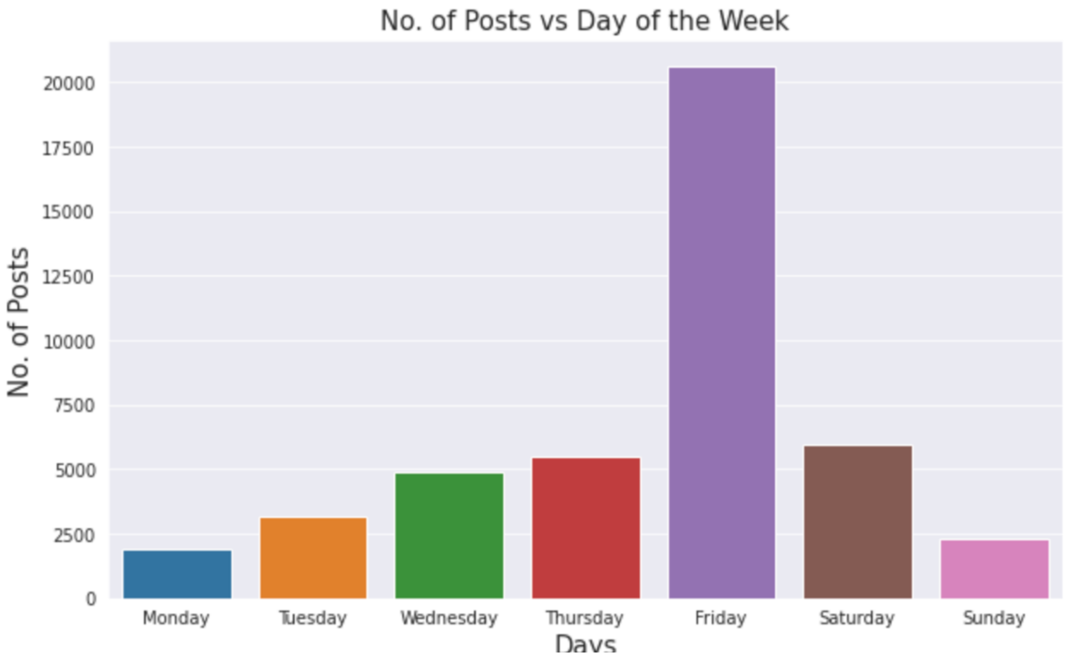


Figure 1: Engagement of reddit users, by day of the week.

1.2 Commonly used words

The frequency analysis chart, figure 2 shows just how dominant the subject of $GME is in the WSB community. This one word had over twice the frequency of the ubiquitous investment term for any stock, ‘buy’.

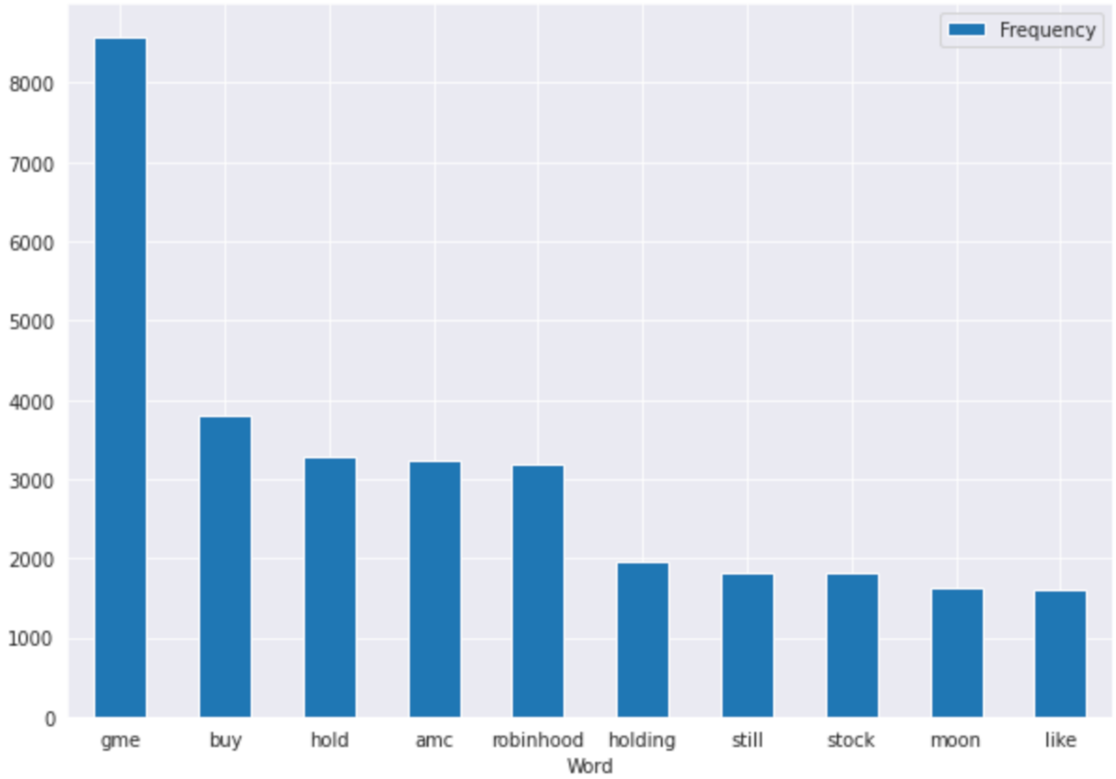


Figure 2: Frequency of occurrence of words related to $GME in the WSB community.

1.3 Emotion Detection

Using the Text2Emotion library, aggregated time series of trends in emotions by day is assembled. Figure 3 shows that 'Fear' is by far the most dominant emotion in the posts we have examined. This supports the hypothesis as fear of missing out, or 'FOMO' is a well-recognised phenomenon that would increase buying if the stock were performing well, whether as fear of substantial losses if the stock is dipping would be an incentive to sell.

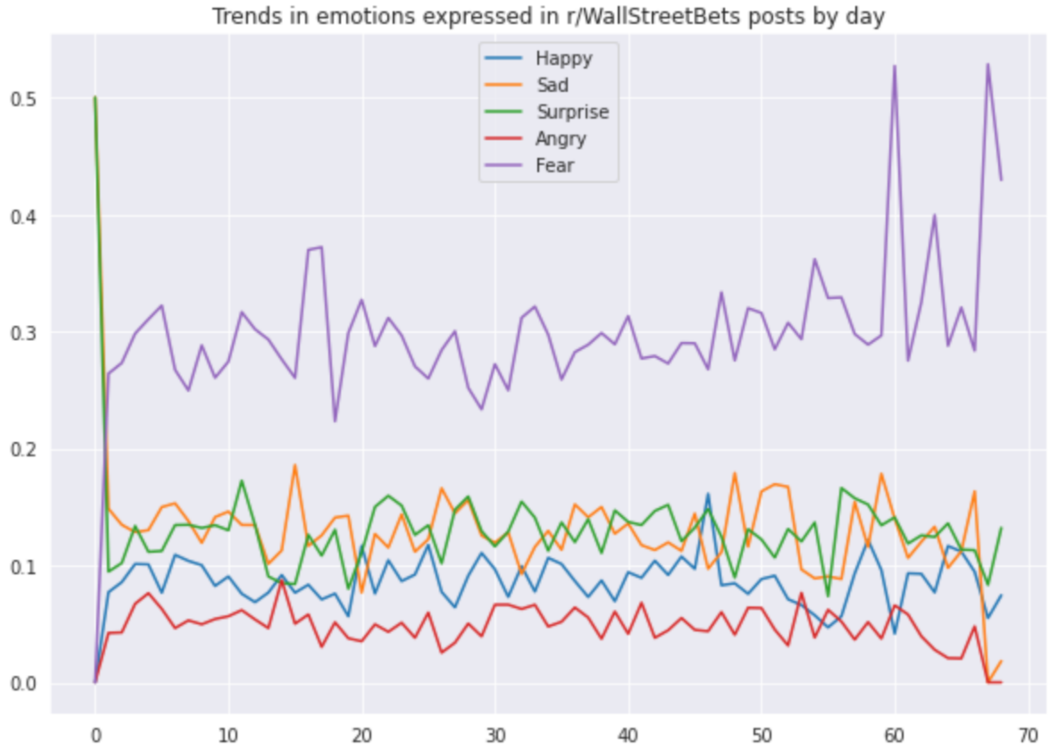


Figure 3: Trend analysis of emotions identified.

1.4 Stock Price

The delta graph (delta is the difference between closing price and opening price) illuminate the volatility of the stock. Time series graph, figure 4 shows the polarity of posts in WSB tended toward positive rather than negative sentiment.

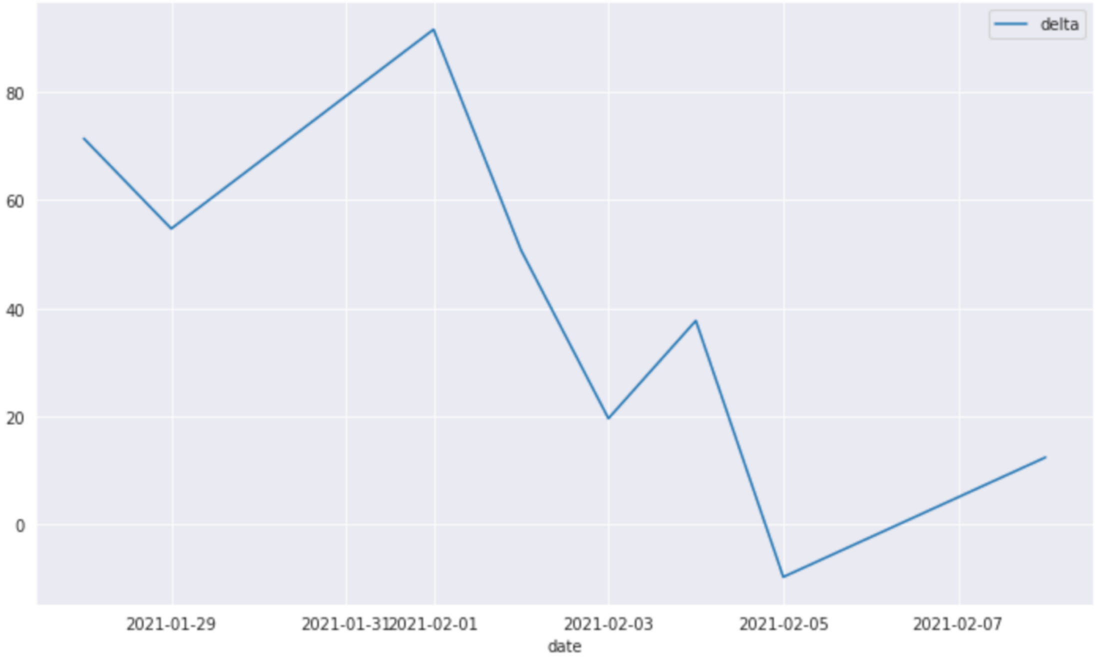
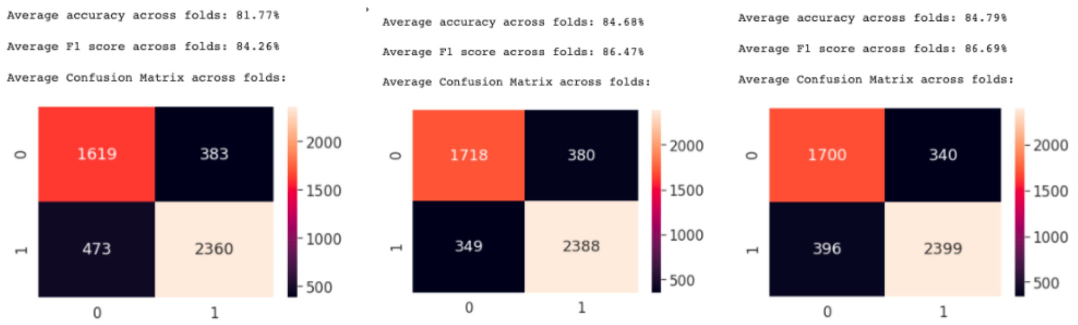


Figure 4: Time Series analysis of $GME stock price.

2 Model Comparison

The interest is in the binary classification of either positive or negative sentiment, as our hypothesis revolves around the core concept of volatility being a result of heightened emotional reaction, hence we have removed neutral labels during data pre-processing. Our three supervised learning models logistic regression, Linear SVC model, and Multinomial Naïve Bayes model have been optimised to predict a post being of one of these two categories. All models were drawn from Scikit-learn's sklearn library.

As shown in figure 5, all performed well in accuracy testing. Multinomial NB model showed a subtle reduction in accuracy when compared to the other 2. The difference in accuracy is negligible and does not make a particularly strong case for use of Linear SVC or Logistic regression.



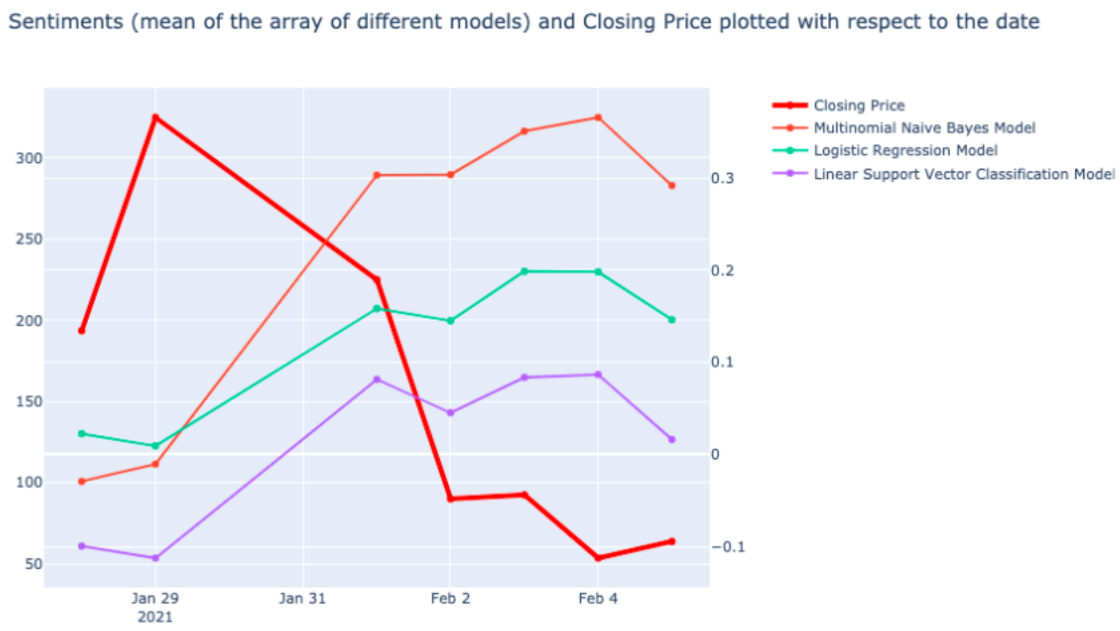
```
Average accuracy across folds: 81.77%
Average F1 score across folds: 84.26%
Average Confusion Matrix across folds:
```

```
Average accuracy across folds: 84.68%
Average F1 score across folds: 86.47%
Average Confusion Matrix across folds:
```

```
Average accuracy across folds: 84.79%
Average F1 score across folds: 86.69%
Average Confusion Matrix across folds:
```

Figure 5: (From left to right): Confusion matrices for Multinomial NB, LinearSVC and Logistic Regression respectively

To produce an easily understandable time series that would highlight the polarity differences between models clearly, we split the mean values by date from Jan 28th to Feb 5th. Figure 6. suggests that the Multinomial NB model was more positive than the LR during the last day (Feb 5th) even though it was more negative in the beginning (Jan 28th). As hypothesised, the LSVC and LR had an almost identical curve and the average tended toward zero(neutral). Although, we cannot conclusively state that any model would provide better performance. This lack of difference suggests that there is much to be improved upon in our data modelling approach.



Sentiments (mean of the array of different models) and Closing Price plotted with respect to the date