Comments	Responses
Give interpretation of Correlation results	We have reflected this
Try step-predictions with PyCaret in place of SARIMA	We utilized SKForecast to turn PyCaret model

FORECASTING ANALYSIS OF LIQUIDITY AND VOLATILITY OF GME STOCK VIA SENTIMENTS FROM REDDIT POSTS

Cansino, Adrian
Torres, Coltrane

Chapter 1
Introduction

Research Questions

- How to explore the relationship between Reddit sentiments and the GME stock variables?
  - How can relevant data (i.e. sentiments, liquidity, and volatility) be prepared?
  - How are Reddit sentiments correlated with the GME stock variables?
  - How can machine learning models (ML) models be implemented for predicting the GME stock variables using Reddit sentiments, historical data, or both?

Research Objectives

- Determine how relevant Reddit posts and GME stock data can be collected and used to prepare Reddit sentiments and the GME stock variables.
- Determine the correlation of Reddit sentiments with the GME stock variables.
- Determine how to implement ML models to use Reddit sentiments, historical data, or both to predict the GME stock variables.

Scope and Limitations

- The study aims to **investigate the association of Reddit with GME**
  - **Only within January 2021 until August 2021**
  - **Only by conducting correlation analysis between Reddit sentiments and the GME stock variables**, and **using ML to make predictions on the GME stock variables using Reddit sentiments.**
    - The GME variables we consider are liquidity and volatility
    - Only text titles from Reddit posts will be used in sentiment analysis
- **Sentiment extraction is assumed to be accurate, further validation is not explored**

Chapter 2
Related Literature

Stock Market

- Liquidity
  - The ratio of volume of shares to the absolute returns over all days with nonzero returns

$$Amivest_i = \sum_{t=1}^{T} Volume_{it} \; / \sum_{t=1}^{T} |return_{it}|$$

- Volatility
  - Standard deviation of stock returns

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(r_i - m)^2}{n-1}}$$

Stock Market Social Media Studies

- HotCopper sentiments and short sales
  - Liquidity can be measured by volume of short sales
  - Positive correlation between negative sentiments and volume of short sales
- StockTwits and Chicago Board of Exchange (CBOE) Volatility Index (VIX)
  - Volatility informs market risk
  - Positive correlation between negative sentiments and perceived market risk

AutoML in Stock Market

- Minimizes the complex processes of machine learning in the financial sector
- Stock price prediction via AutoML
  - Regression models and neural networks were used to make 1-day and 10-day predictions
    - Favorable evaluation scores in 1-day prediction; predicting precise values is best suited for short time windows
    - However, predicting trends is still viable in either case

Time Series Forecasting Studies

- Incidence Forecasting of Schistosomiasis in China
  - A low MSE score was found, indicating a good performance of the SARIMA model.
- Previous studies found that time series forecasting models (e.g., ARIMA, SARIMA, etc.) performs reasonably well in short term forecasting.

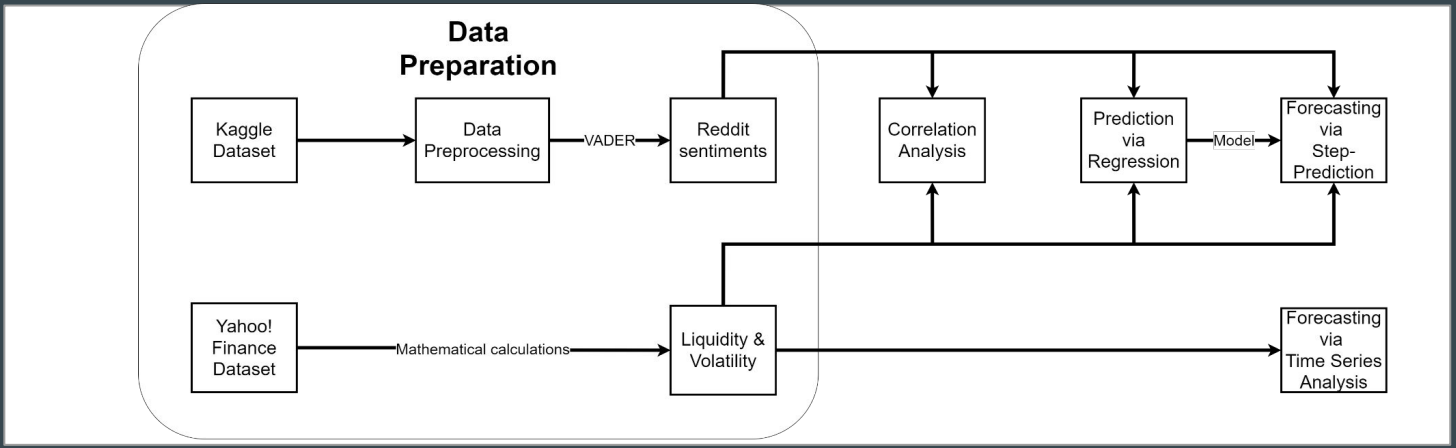
Similarities

- **Extraction of sentiments from social media data** (e.g., tweets) was conducted, then **sentiments were used as stock return predictive features** for ML models.
- **Sentiments and extracted values of stock variables (e.g., liquidity, volatility) were used as data for correlation analysis** - to identify the influence of social media sentiments to the stock variables.
- **Stock price data were collected, and used as predictive features** to be fed in machine learning models, both in AutoML and time series studies.

Chapter 3
Methodology

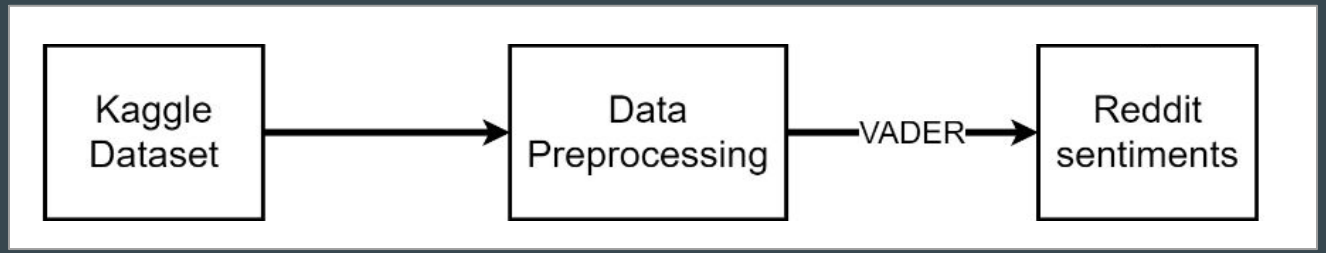
Overview

- Data preparation
  - Reddit dataset from Kaggle
  - GME dataset from Yahoo! Finance
- Correlation analysis between sentiments and stock variables
- Regression prediction and step-prediction incorporating Reddit sentiments
- Time Series Forecasting



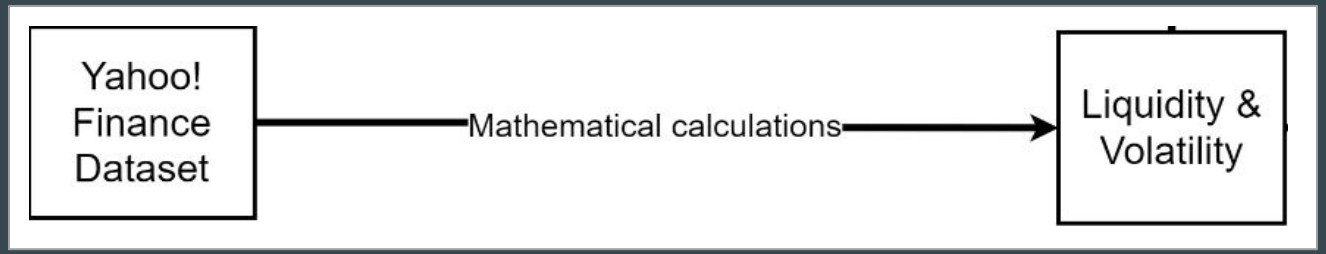
Data Preparation

- Data Preprocessing
  - Filter GME Posts from Reddit dataset
    - Filter keywords: gamestop, gme
    - Columns used: timestamp and title
  - Preprocess text
    - Remove duplicate spaces and stopwords, sentence tokenization
- Sentiments extraction with VADER
  - Resulting sentiments aggregated daily by mean



Data Preparation

- Liquidity and Volatility extracted via mathematical formulae in related literature



Correlation Analysis

- Pearson: `scipy's pearsonr()`
- Cross-correlation: `numpy's correlate()`
  - Correlation between variables at different times
  - Insights like: leading variable, time window when correlation is strongest

Regression with PyCaret

- PyCaret trains regression models using Reddit sentiments
  - Three feature sets considered in the setup
    - Compound and polar sentiments
    - Compound sentiments
    - Polar sentiments
- Default configurations were used (ex. 10-fold cross validation, 70-30 random train-test split)
- Three setups
  - Best models acquired by Pycaret's `compare_models()`
  - Best model overall acquired by comparing the best model in each setup

Regression Models in PyCaret

- PyCaret's regression module offers several regression models

Model	Abbreviation
Linear Regression	lr
Lasso Regression	lasso
Ridge Regression	ridge
Elastic Net	en
Least Angle Regression	lar
Lasso Least Angle Regression	llar
Orthogonal Matching Pursuit	omp
Bayesian Ridge	br
Automatic Relevance Determination	ard
Passive Aggressive Regressor	par
Random Sample Consensus	rsc
TheilSen Regressor	tsr
Huber Regressor	huber
Kernel Ridge	kr
Support Vector Machine	svm
K Neighbors Regressor	knn
Decision Tree	dt
Random Forest	rf
Extra Trees Regressor	et
AdaBoost Regressor	ada
Gradient Boosting Regressor	gbr
Multi Level Perceptron	mlp
Extreme Gradient Boosting	xgboost
Light Gradient Boosting	lightgbm
CatBoost Regressor	cbr

Step-prediction with SKForecast

- SKForecast makes step-predictors using Reddit sentiments and historical data for predictions
  - Like PyCaret, the same three feature sets were considered
- `regressor` parameter uses PyCaret final models
- `lags` parameter tuned by grid search over values of 1 to 20
- Like PyCaret, three setups
  - Best models selected after tuning their `lags` parameter
  - Best model overall selected after comparing the best model in each setup

Time Series Analysis with SARIMA

(1) Stationarity checks
   - Rolling Statistics & ADF
   - Differencing

(2) Hyperparameter tuning

(3) Forecast of Liquidity and Volatility

Chapter 4
Results

Correlation

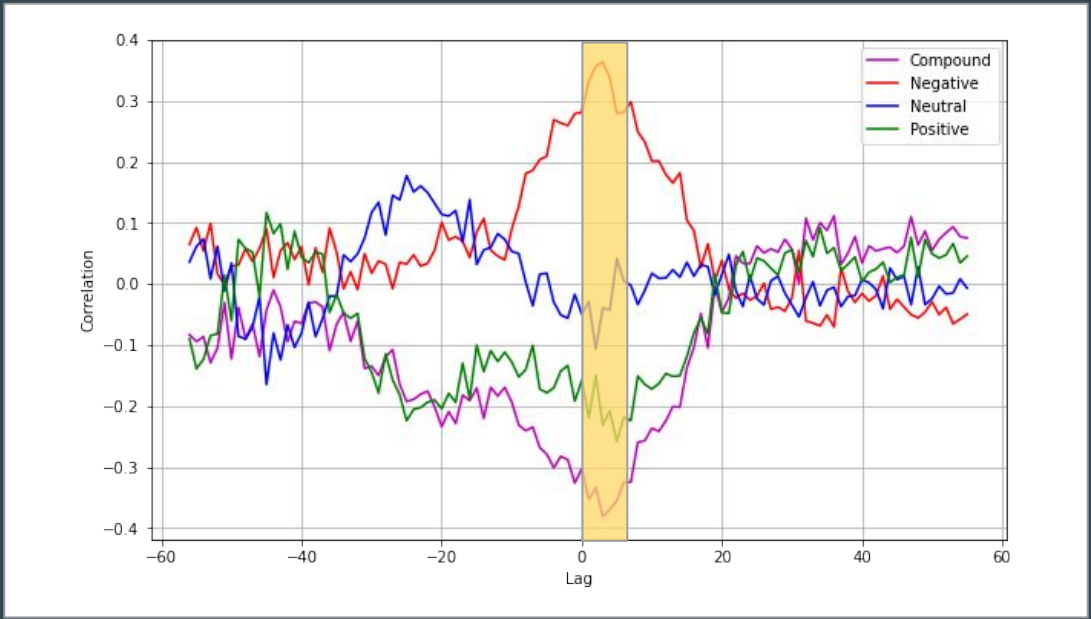
Pearson

- Neutral Sentiments: low correlation
- Strength of correlations with stock variables (from strongest to weakest): Compound, Negative, Positive and Neutral
- Compound, Neutral, and Positive: positive correlation with either of the stock variables
- Negative: positive correlation for both stock variables

Sentiment	Liquidity	Volatility
Compound	-0.303	-0.405
Negative	0.281	0.341
Neutral	-0.049	-0.022
Positive	-0.158	-0.234

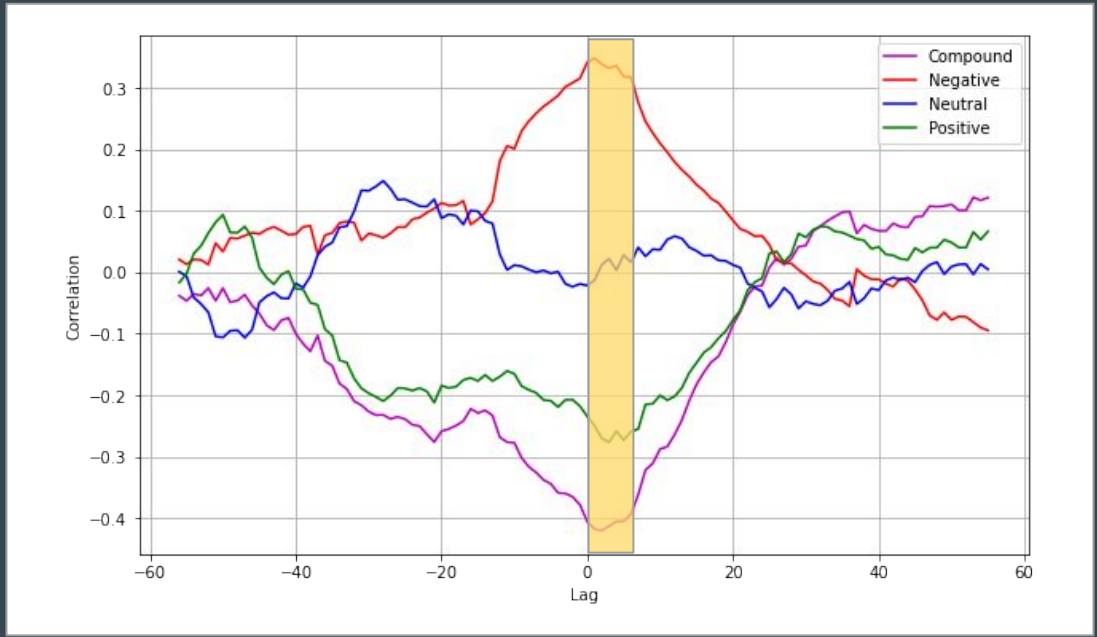
Cross-Correlation (Liquidity)

- Strongest correlation values generally hovering around 0 lag
- These values can generally be found at positive lags ⇒ sentiment leads liquidity



Cross-Correlation (Volatility)

- Similar observations as the cross-correlation between liquidity and sentiments can also be observed in the cross-correlation between volatility and sentiments



Cross-Correlation

- Strongest correlations generally at small positive lags ⇒ **sentiments lead target variables short term**
  - i.e. Compound sentiments from 3 days past have strongest relationship with present liquidity and so on
  - Neutral sentiments show outliers of -25 and -28 lags
- Order of correlation strength is consistent with Pearson

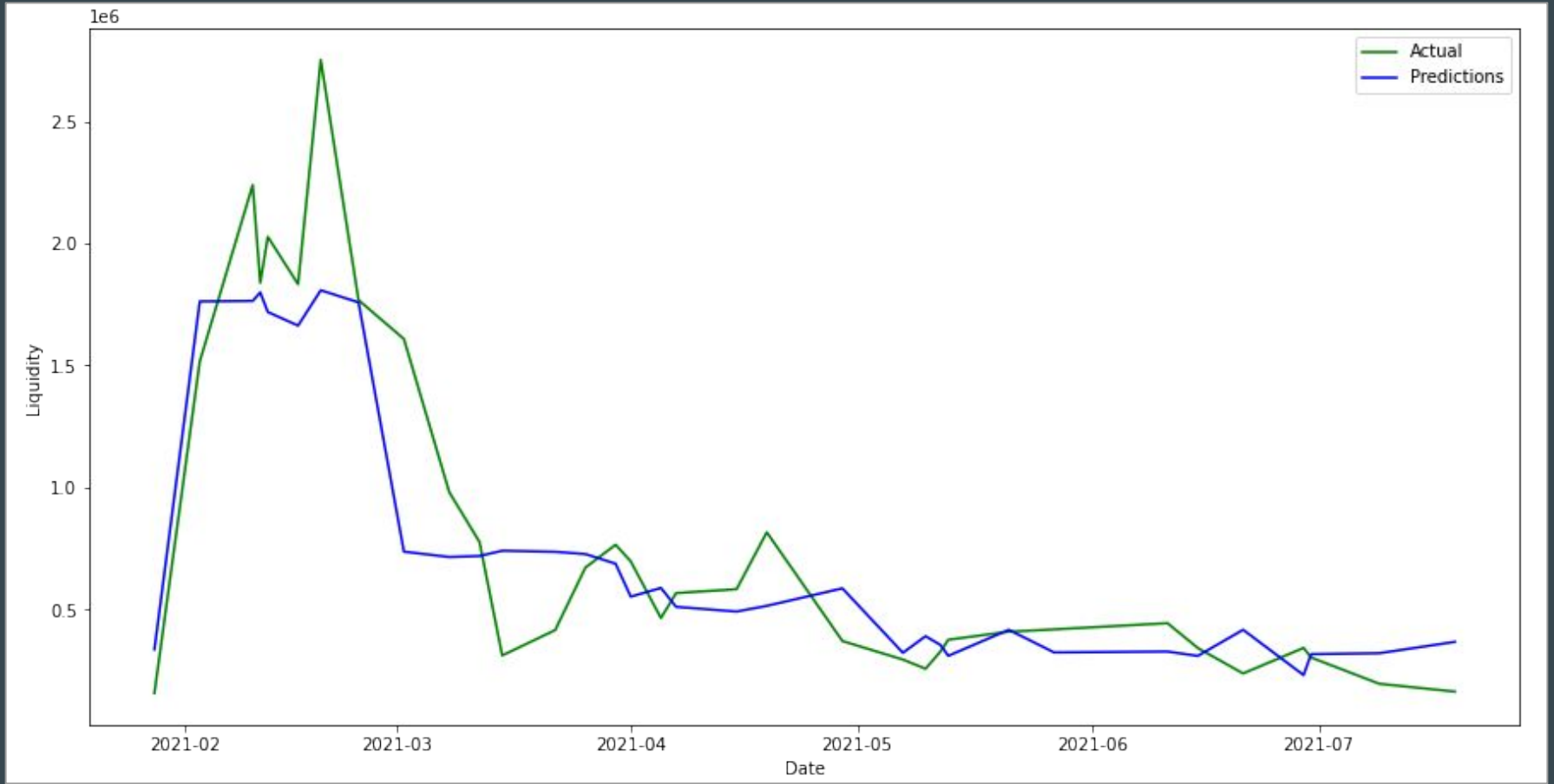
Sentiment	Liquidity		Volatility	
	Lag	Correlation	Lag	Correlation
Compound	**3**	**-0.381**	**2**	**-0.420**
Negative	**3**	**0.364**	**1**	**0.348**
Neutral	-25	0.177	-28	0.149
Positive	**5**	**-0.258**	**3**	**-0.277**

Predictions

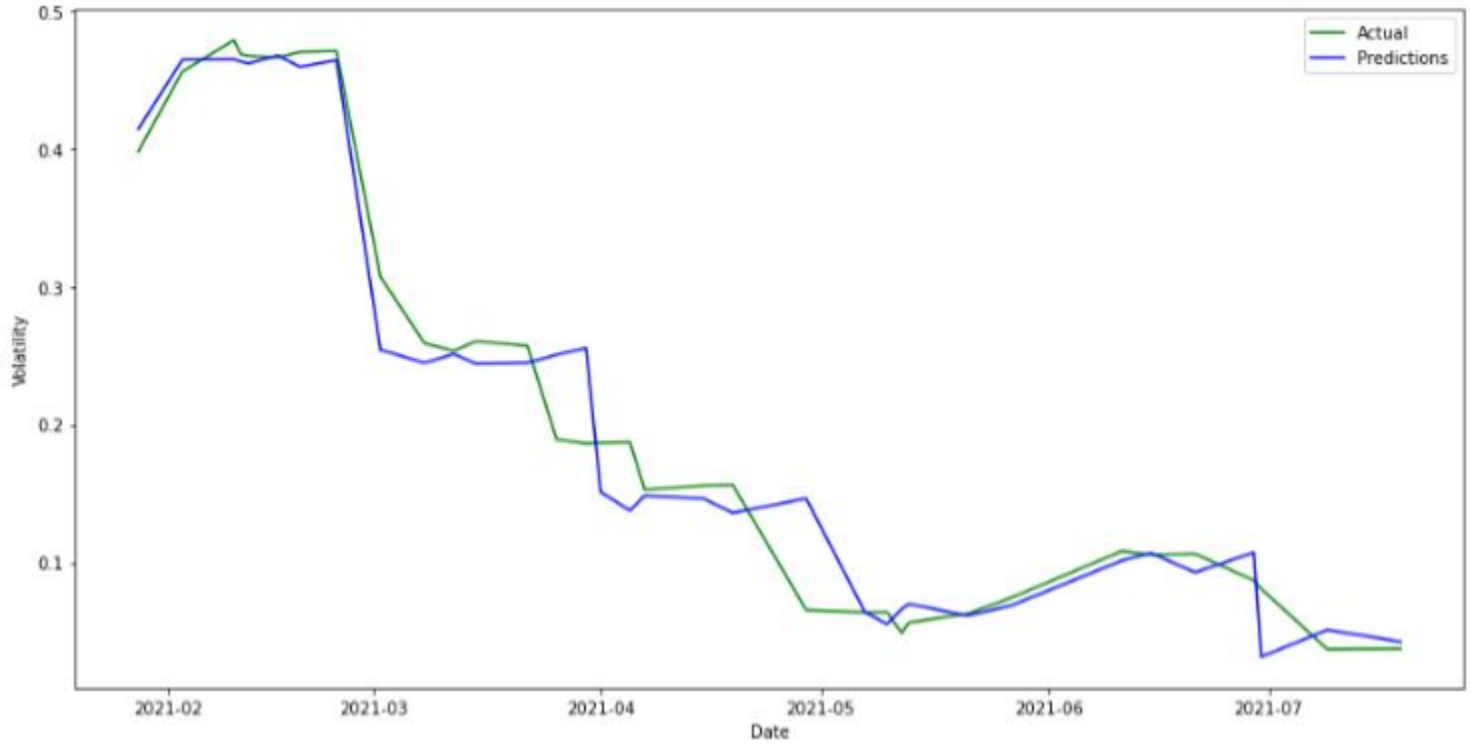
PyCaret Models Comparisons

Model	All Sentiments		Compound Sentiments		Polar Sentiments	
	Liquidity	Volatility	Liquidity	Volatility	Liquidity	Volatility
ada	384170.1056	0.0450	395173.0483	0.0506	**386282.5098**	0.0425
br	590043.6941	0.0393	590043.6941	0.0384	590043.6941	0.0399
dt	519659.4885	0.0504	612838.4052	0.0462	508512.6965	0.0490
en	639251.4016	0.1410	738105.4766	0.1410	640072.7148	0.1410
et	529794.4156	0.0588	528048.2479	0.0607	561673.5923	0.0591
gbr	457278.0421	0.0587	456295.4093	0.0534	455897.2637	0.0587
huber	594939.5441	**0.0373**	594943.4782	**0.0372**	594974.8171	**0.0372**
knn	480990.6312	0.0809	480748.1906	0.0749	436310.5263	0.0784
lar	2842325.9287	0.1410	433027.9529	0.1410	432879.6588	0.1410
lasso	433071.9938	0.1410	433719.1328	0.1410	432131.9719	0.1410
lightgbm	531585.8842	0.1347	539964.1700	0.1328	531799.4380	0.1328
llar	432957.8873	0.0426	433020.7801	0.0408	432183.1723	0.0412
lr	437078.0812	369.1386	433492.7016	23.1609	1423658621.4547	154.2409
omp	**356573.8006**	0.0404	**381183.9570**	0.0419	408589.1437	0.0403
par	602836.0572	0.0739	574709.1468	0.0742	602235.4540	0.0733
rf	518822.3724	0.0565	520972.3928	0.0591	538664.4192	0.0563
ridge	431206.3922	0.0407	445972.8039	0.0415	430939.2031	0.0405
xgboost	561450.4297	0.0449	575351.0242	0.0454	563042.7031	0.0448

OMP Predictions on Liquidity



Huber Predictions on Volatility



Final PyCaret Models Performance

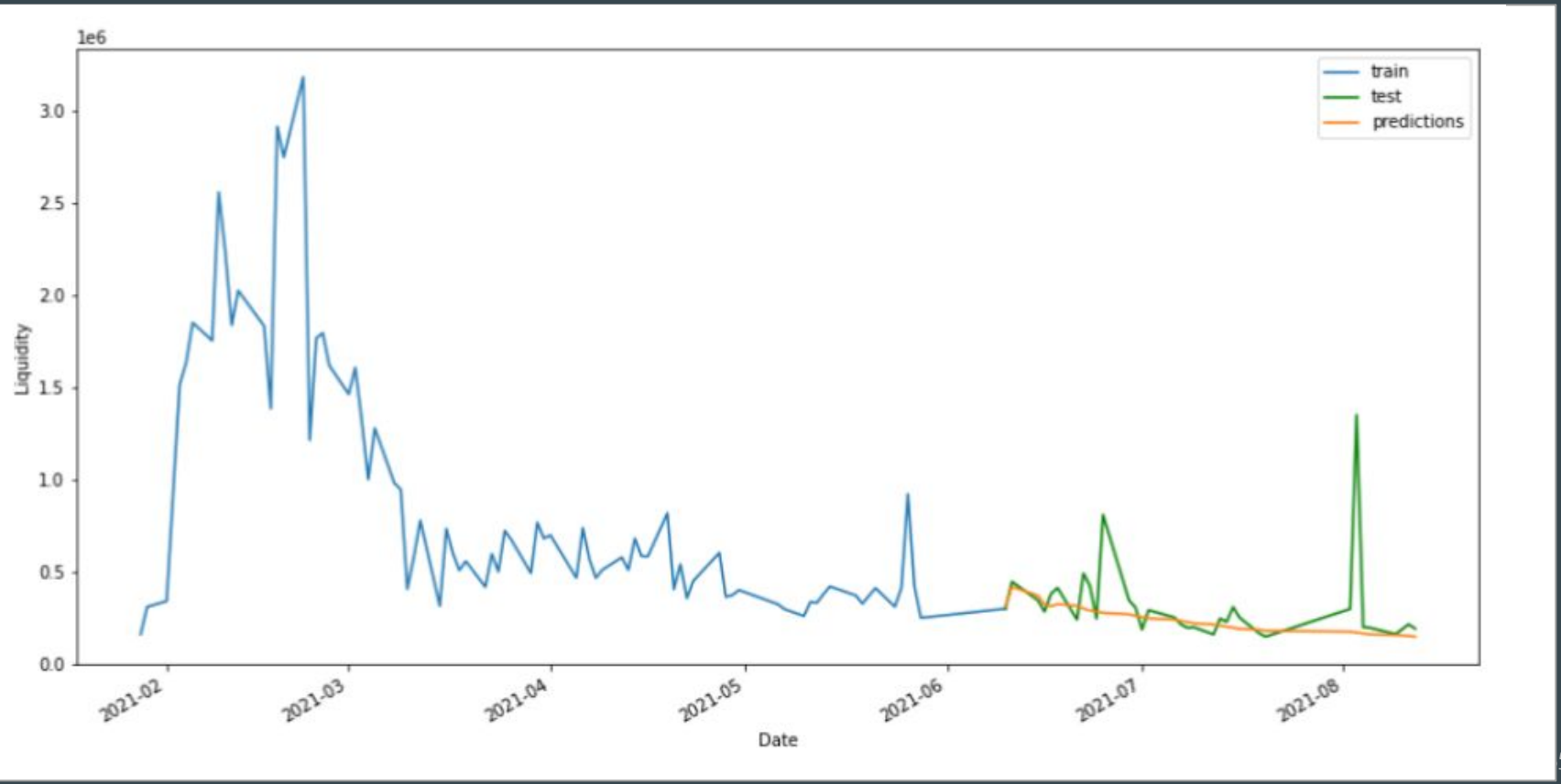
	Train	Test
Liquidity	0.117	0.094
Volatility	0.083	0.063

Forecast

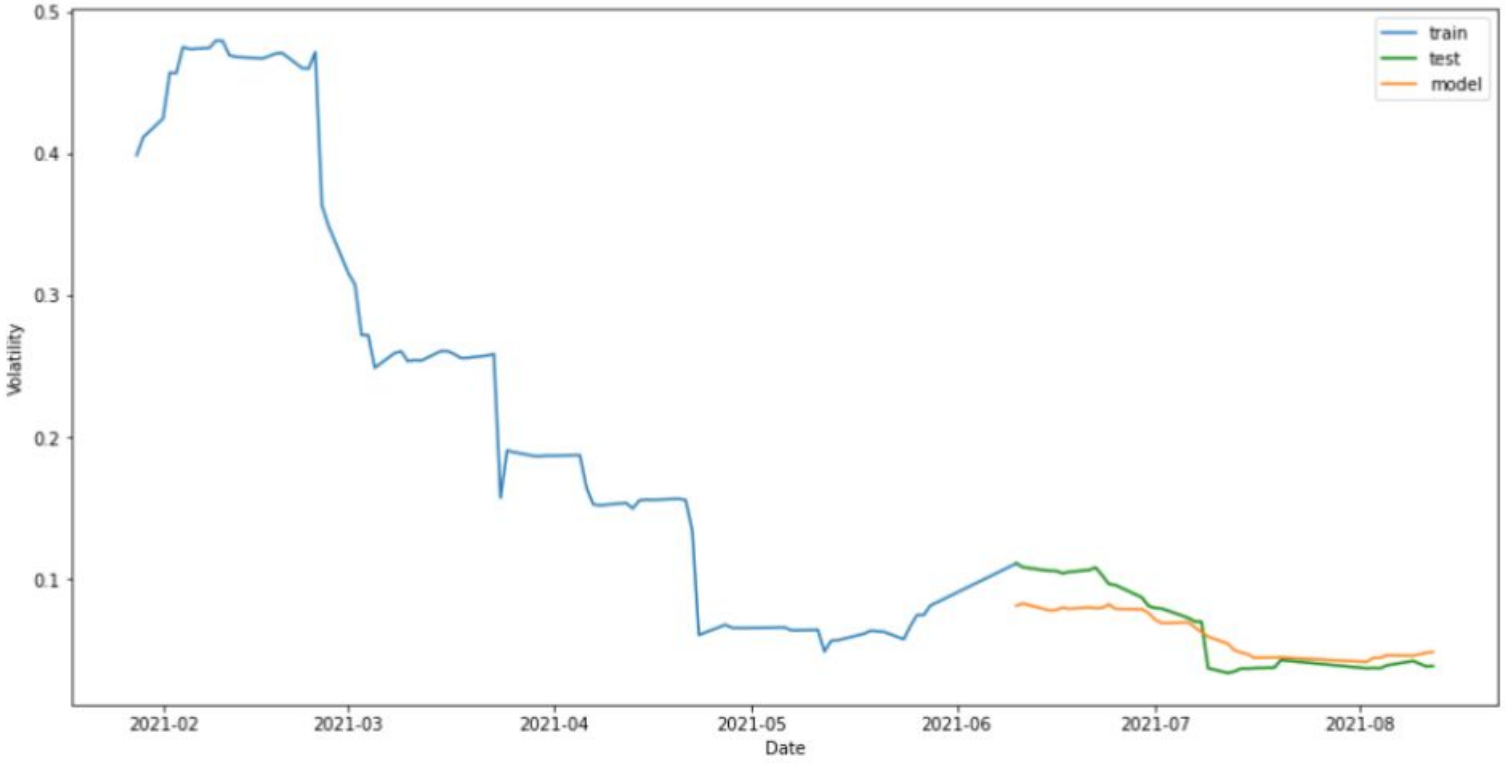
SKForecast Models Comparison

Lag	All Sentiments		Compound Sentiments		Polar Sentiments	
	Liquidity	Volatility	Liquidity	Volatility	Liquidity	Volatility
1	0.1647	0.0460	-	0.0519	-	**0.0462**
2	0.1121	0.0449	-	0.0485	-	0.0468
3	0.1534	0.0439	-	0.0422	-	0.0487
4	0.0776	0.0451	-	**0.0396**	-	0.0532
5	0.0774	0.0667	-	0.0488	-	0.0746
6	0.0783	0.0594	-	0.0496	-	0.0750
7	**0.0773**	0.0712	-	0.0609	-	0.0896
8	0.0822	0.0659	-	0.0465	-	0.0701
9	0.0886	0.0494	-	0.0398	-	0.0578
10	0.0817	0.0723	-	0.0608	-	0.0884
11	0.0818	0.0778	-	0.0636	-	0.0888
12	0.0883	0.0664	-	0.0548	-	0.0812
13	0.0887	0.0411	-	0.0400	-	0.0677
14	0.0892	0.0394	-	0.0418	-	0.0802
15	0.0889	**0.0372**	-	0.0420	-	0.0587
16	0.0947	0.0413	-	0.0442	-	0.0597
17	0.0913	0.0439	-	0.0494	-	0.0556
18	0.0902	0.0695	-	0.0779	-	0.0665
19	0.0961	0.0976	-	0.1135	-	0.1009
20	0.0975	0.1095	-	0.1236	-	0.1130

OMP Step-predictor Forecasts on Liquidity



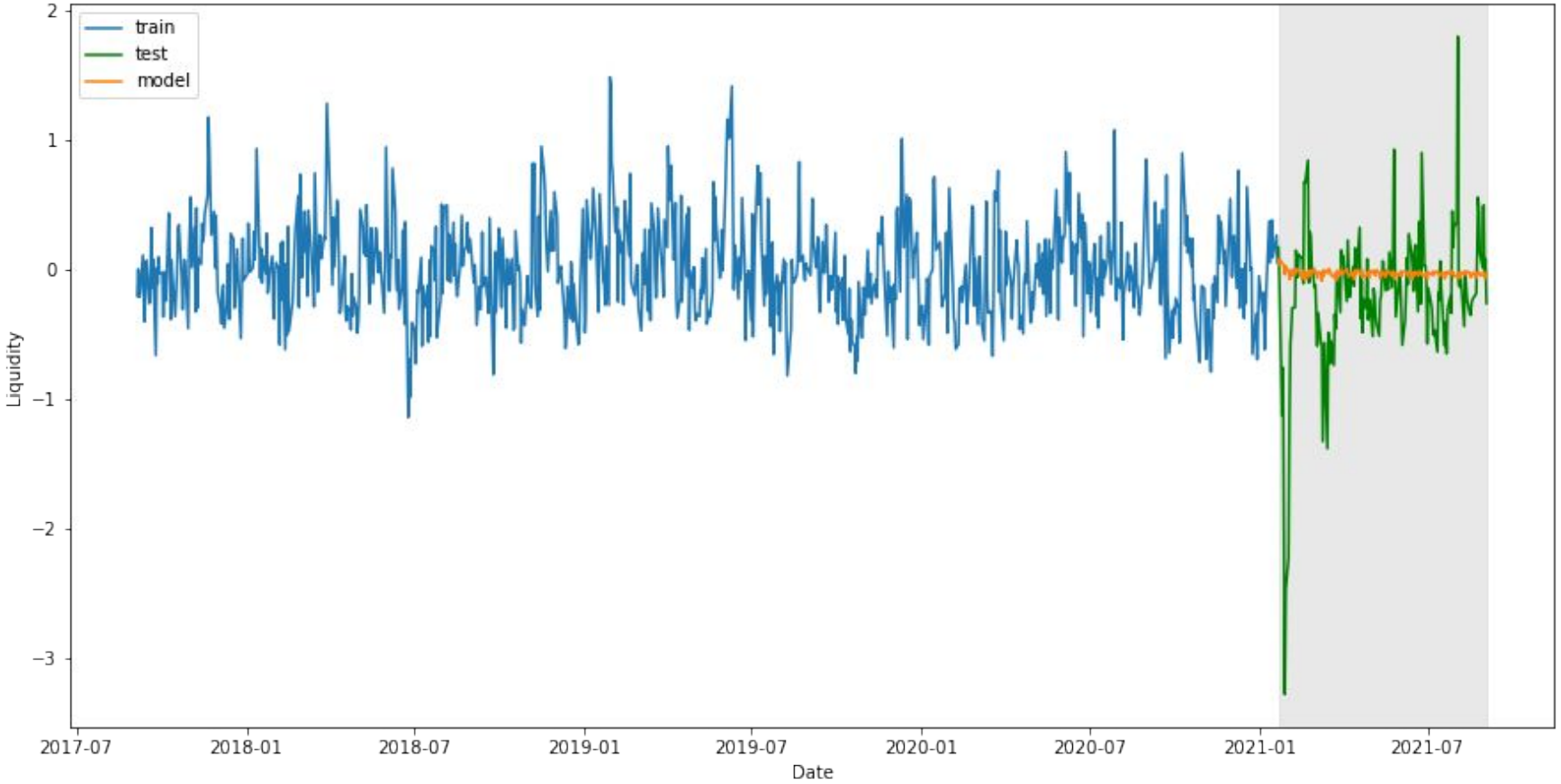
Huber Step-predictor Forecasts on Volatility



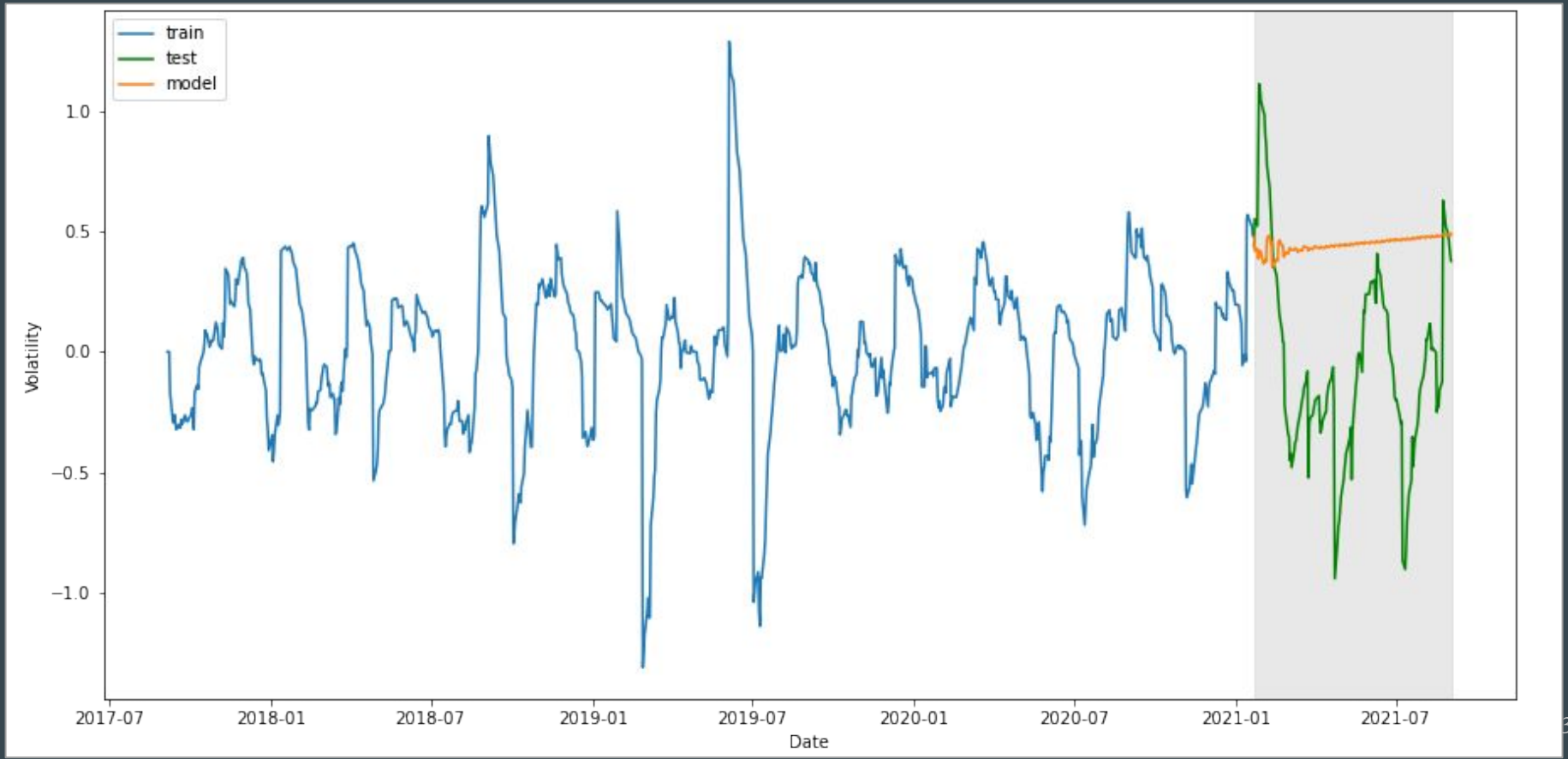
SARIMA Model Performance

	Train	Test
Liquidity	0.066	0.114
Volatility	0.043	0.249

SARIMA Predictions on Liquidity



SARIMA Predictions on Volatility



Chapter 5
Conclusions & Recommendations

Conclusions

RQs	Conclusions
How can relevant data (i.e. sentiments, liquidity, and volatility) be prepared?	- Reddit dataset was collected from Kaggle - Reddit sentiments were extracted via VADER - GME stock dataset was collected from Yahoo! Finance - GME liquidity and volatility were extracted via mathematical formulas
How are Reddit sentiments correlated with the GME stock variables?	- **Significant correlations were found within short time windows, with sentiments generally leading stock variables**
How can machine learning (ML) models be implemented for predicting the GME stock variables using Reddit sentiments, historical data or both?	In predicting the GME stock variables: - **Regression via PyCaret** proved viability of using **Reddit sentiments** - **Step-prediction via SKForecast** proved viability of using both **Reddit sentiments and historical data** - **Time series analysis via SARIMA** proved comparably worse than the others using only **historical data**

Recommendations

- Future research may benefit from exploring
  - other NLP libraries in extracting sentiments
  - validation methods for sentiments
  - other additional hyperparameters
  - other optimization techniques for tuning hyperparameters

22nd Philippine Computing Science Congress (PCSC 2022),
Virtual Conference, Philippines
Computing Society of the Philippines

Correlation and Prediction of Gamestop Stock Liquidity and Volatility using Reddit Sentiments

Adrian Joshua Cansino, Coltrane Torres and Jann Railey Montalan
Ateneo de Manila University, Quezon City

Abstract

In January 2021, GameStop (GME) stock experienced a drastic increase in stock prices. This phenomena, dubbed GME Short Squeeze, is purportedly driven by Reddit community r/wallstreetbets. This study aims to investigate whether Reddit's influence on GME stocks were actually significant by finding the correlation between Reddit sentiments and GME stock liquidity and volatility, and employing machine learning (ML) models to determine if Reddit sentiments can be used as predictive features of GME stock liquidity and volatility.

Calculating the Pearson Correlation and cross-correlation, the study finds a correlation between the variables, and that the trend of Reddit sentiments predates similar trends in GME stock liquidity and volatility. Moreover, using PyCaret, the study also finds that Orthogonal Matching Pursuit and Huber Regressor to be the best performing model for predicting liquidity and volatility using sentiments, respectively. Furthermore, their reasonably low Root Mean Squared Error (RMSE) of the models suggest a decent performance.

Objectives

The goal of the study is to investigate the influence of Reddit on GME stocks by conducting correlation analysis between Reddit sentiments and GME stock liquidity and volatility, and AutoML to determine if Reddit sentiments can be used as predictive features of GME stock liquidity and volatility.

Methodology

Data Preparation
A Reddit dataset containing Reddit posts from r/wallstreetbets spanning from January 2021 to August 2021 was retrieved from Kaggle, while a dataset containing historical GME stock data that covers the same time period was pulled from Yahoo!Finance. After data collection, sentiments from Reddit posts in r/wallstreetbets, and stock liquidity and volatility from past GME stock data were extracted.

Extraction of Sentiments from Reddit Data. In sentiment extraction, the Reddit dataset was first filtered to include only GME-related posts using keywords like gme, gamestop, and $gme. Then, only the relevant columns (timestamp and title) were kept. Afterwards, stop words were removed using the Natural Language Toolkit's built-in stop words removal. URLs and duplicated white spaces were also removed. Lastly, sentence tokenization was performed.

After preprocessing, sentiments were extracted following a rule-based approach with the pre-trained lexicon Valence Aware Dictionary for Sentiment Reasoning (VADER). VADER is commonly used due to its sensitivity to both the polarity and emotion intensity of texts. Here, VADER was used to calculate the sentiment ratings of entire texts, which were then aggregated by mean to match the daily stock data.

Extraction of Stock Liquidity and Volatility from Stock Data. To extract liquidity from the stock data, the ratio of the volume of shares traded to the absolute returns over all days with nonzero returns is taken [1].

$$Amivest_l = \sum_{t=1}^{T} Volume_{it} \Big/ \sum_{t=1}^{T} |return_{it}|$$

On the other hand, volatility is extracted from the stock data by taking the standard deviation of stock returns [3].

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (r_i - m)^2}{n-1}}$$

Correlation Analysis of Sentiments and Stock Liquidity and Volatility
Analysis of the Pearson Correlation and cross-correlation of sentiments with stock liquidity and volatility was done to investigate their relationships.

The Pearson Correlation gives the general correlation between the variables and was done via scipy's pearsonr.

Cross-correlation gives the correlation over time, providing information such as which variable leads the relationship, or at what time lag or difference is their correlation strongest. Cross-correlation was done via numpy's correlate after normalizing the data to have normal correlation values.

Prediction using Machine Learning models via PyCaret
PyCaret is a Python AutoML library that was used to conduct AutoML for predicting stock liquidity and volatility using sentiments. PyCaret's regression module was used to determine the best regression model for predicting the stock variables based on the RMSE, a statistical metric for measuring regression model performance via its error. Given that the RMSE describes the average raw error of the model, it was normalized by dividing the range of the data in order to have normalized RMSE results.

Results

Correlations between Sentiment and Stock Liquidity and Volatility
The Pearson Correlation showed that compound sentiments had the strongest correlation with both liquidity and volatility, followed by negative, positive, and neutral sentiments. Moreover, Table 1 shows that negative sentiments had a positive correlation with liquidity and volatility, whereas compound, neutral, and positive sentiments showed the opposite, showing consistency with previous studies [2,4].

Sentiment	Liquidity	Volatility
Compound	-0.303	-0.405
Negative	0.281	0.341
Neutral	-0.049	-0.022
Positive	-0.158	-0.234

Table 1. Pearson Correlation Values between Sentiments, and Liquidity and Volatility

Afterwards, cross-correlation elucidated the relationship between the target variables and sentiments in greater detail. Except for neutral sentiments, the strongest correlation values between the remaining sentiments and both of the variables appeared in the middle towards a lag of zero as shown in Figures 1 and 2. Moreover, these points seem closer to a positive lag. That the strongest correlations appear at small positive lags imply that 1) the correlation is stronger with shorter lags, and 2) the connection is strongest when sentiments lead liquidity and volatility.

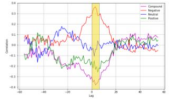


Figure 1. Cross-Correlation of Liquidity with Sentiments

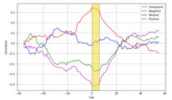


Figure 2. Cross-Correlation of Volatility with Sentiments

With Table 2, it is evident that both liquidity and volatility were found to be positively correlated with negative sentiments in both correlations, as well as being negatively correlated with positive and compound sentiments. This is consistent with previous research that found links between the stock variables and sentiments [2,4]. This suggests that likewise, volatility is driven by low confidence or fear in the market, and that liquidity is driven by shorting, which is also associated with low confidence. Moreover, cross-correlation uncovered that unlike neutral sentiments, compound, positive and negative sentiments from shorter time period in the past were more significantly correlated with present values of the stock variables.

Sentiment	Liquidity		Volatility	
	Lag	Correlation	Lag	Correlation
Compound	3	**-0.381**	2	**-0.420**
Negative	3	**0.364**	1	**0.348**
Neutral	-25	0.177	-28	0.149
Positive	5	**-0.258**	3	**-0.277**

Table 2. Strongest Cross-Correlation Values and Lag between Sentiments, and Liquidity and Volatility

Prediction of Stock Liquidity and Volatility using Sentiments
PyCaret was used to automatically select and train the regression models with the best RMSE in performing regression on sentiments. As Table 3 shows, PyCaret discovered that the Orthogonal Matching Pursuit model was the best model for predicting liquidity with sentiments. On the other hand, PyCaret demonstrated that the Huber Regressor was the best model for predicting volatility.

	Train	Test
Liquidity (Orthogonal Matching Pursuit)	0.117	0.094
Volatility (Huber Regressor)	0.084	0.064

Table 3. RMSE of Final PyCaret Models

Furthermore, the PyCaret models were able to predict market liquidity and volatility patterns with sentiments as seen in Figures 3 and 4, implying that Reddit sentiments may be utilized to predict GME stock data to some extent.

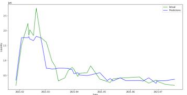


Figure 3. PyCaret Model Predictions on Liquidity

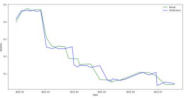


Figure 4. PyCaret Model Predictions on Volatility

Conclusion

This study aimed to find whether Reddit actually had a significant influence on the GME stock by investigating the relationship of Reddit sentiments and stock variables, as well as conducting ML predictions on the stock variables using Reddit sentiments.
Firstly, Reddit data was collected from Kaggle, and GME stock data from Yahoo! Finance. Afterwards, data was preprocessed in a variety of ways to prepare it for the extraction of relevant data. Finally, Reddit sentiments were extracted via VADER, while liquidity and volatility was extracted via mathematical calculations.
Afterwards, correlation analysis on Pearson Correlation and cross-correlation was conducted to investigate the relationship between sentiments and the stock variables, wherein it was found that the stock variables have a strong correlation with compound, as well as negative and positive sentiments at shorter time frames. Moreover, sentiments were found to lead both stock variables.
Finally, the prediction of GME stock liquidity and volatility using sentiments from Reddit posts via PyCaret discovers that the low RMSE of the models selected by PyCaret demonstrate the viability of using Reddit sentiments as predictive features for predicting stock trends.

References

[1] Hyuk Choe and Cheol-Won Yang. 2008. Comparisons of Liquidity Measures in the Stock Markets. (2008), 1767–1822.

[2] Tianyou Hu and Arvind Tripathi. 2015. The Effect of Social Media on Market Liquidity. Available at SSRN 2661099 (2015).

[3] D Mamtha and K Sakthi Srinivasan. 2016. Stock market volatility–conceptual perspective through literature survey. Mediterranean Journal of Social Sciences 7, 1 (2016), 208.

[4] Juan Piñeiro-Chousa, Marcos Vizcaíno-González, and Ada María Pérez-Pico. 2017. Influence of social media over the stock market. Psychology & Marketing 34, 1 (2017), 101–108.

THANK YOU