Honors Thesis: Predicting Bitcoin Price Trend using Sentiment Analysis

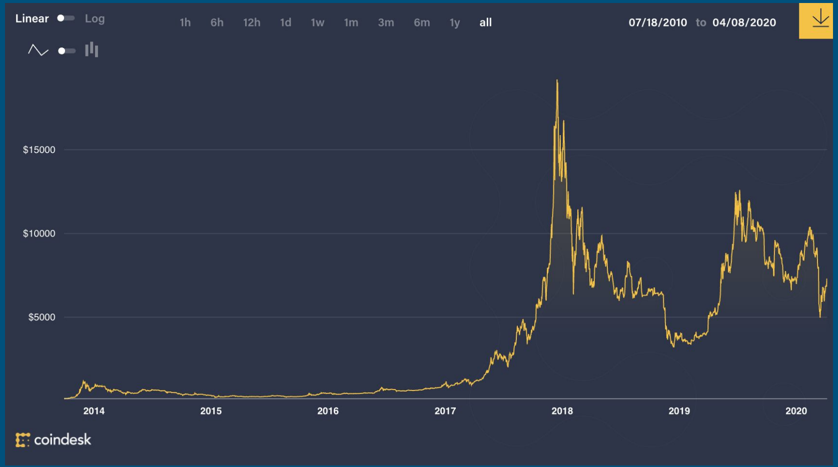
Sam Steinberg
ASU Barrett Honors College
4/8/20

Agenda

- Problem Definition
- Solution
- Steps
- Results
- Future Work

Problem Definition

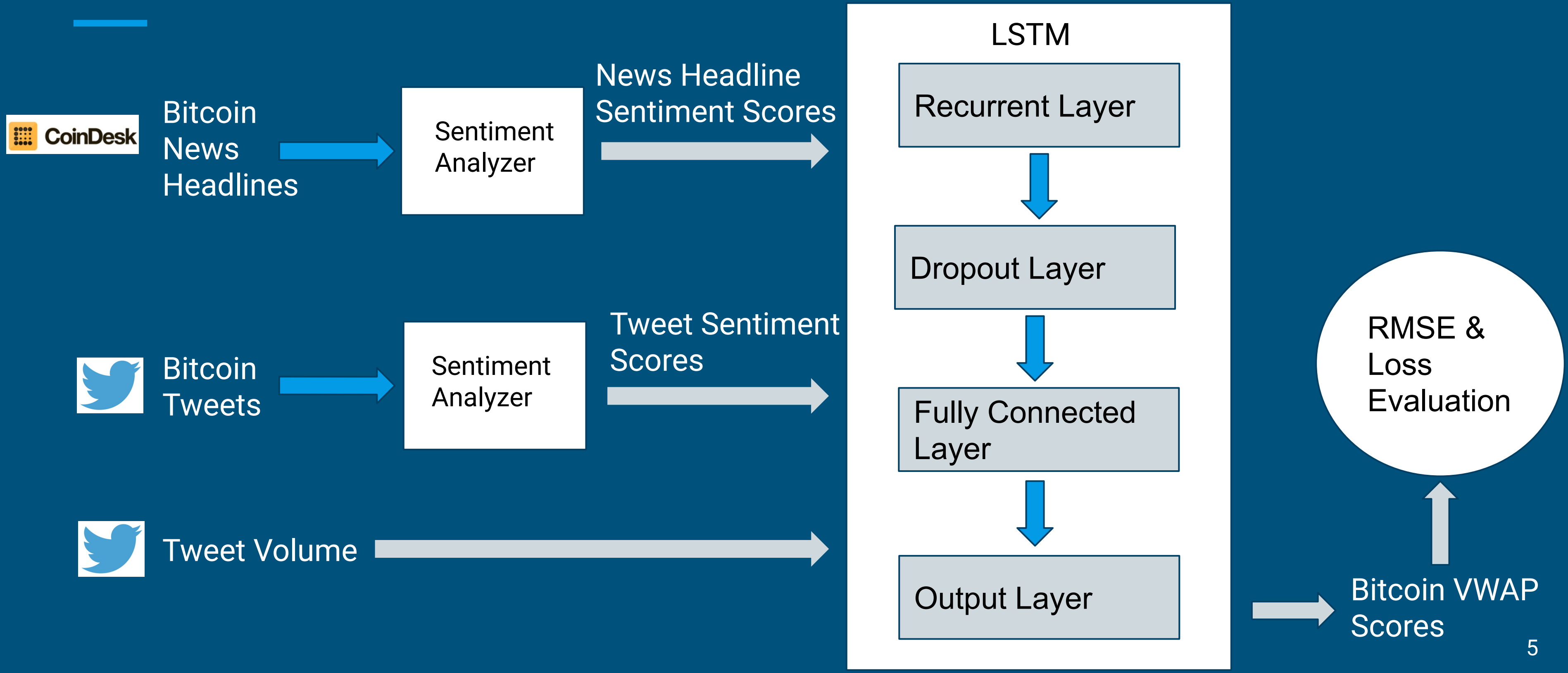
- Unlike some stocks and other investments, Bitcoin is highly volatile and difficult to predict
  - Dec. 2017- 20,000+
  - Dec. 2018- < 4,000
- No effective methods for doing so
- **Goal:** create a model that can accurately predict the price direction of Bitcoin



Solution

- Sentiment Analysis
    - Classifies text as either positive (1), neutral (0), or negative (-1)
    - Top hedge funds use this to influence their investments
    - Analyse both user data and news data
        - People are reactionary to breaking news
- Twitter
    - Mass amounts of user data on Bitcoin
    - Retrieve tweets that include “Bitcoin” or “BTCUSD”
- Model- Long Short-Term Memory (LSTM) Neural Network

Solution


CoinDesk
Bitcoin News Headlines
Sentiment Analyzer
News Headline Sentiment Scores
Bitcoin Tweets
Sentiment Analyzer
Tweet Sentiment Scores
Tweet Volume
LSTM
Recurrent Layer
Dropout Layer
Fully Connected Layer
Output Layer
Bitcoin VWAP Scores
RMSE & Loss Evaluation

Step 1- Preliminary Decisions

- Language- Python
    - Comfortable language
    - Supported, detailed libraries for neural networks
- Neural Network- LSTM
    - Dealing with non-linear, time-series data
    - Accurate even with data gaps [2]
    - Powerful “memory”
- Output: Volume-Weighted Average Price (VWAP) Score (3.25 hour time interval)
    - VWAP- good indicator of a crypto’s market
    - Bitcoin is highly volatile, difficult to predict price trend by month, week
    - Sample size is too small at smaller intervals

Step 2- Data Collection

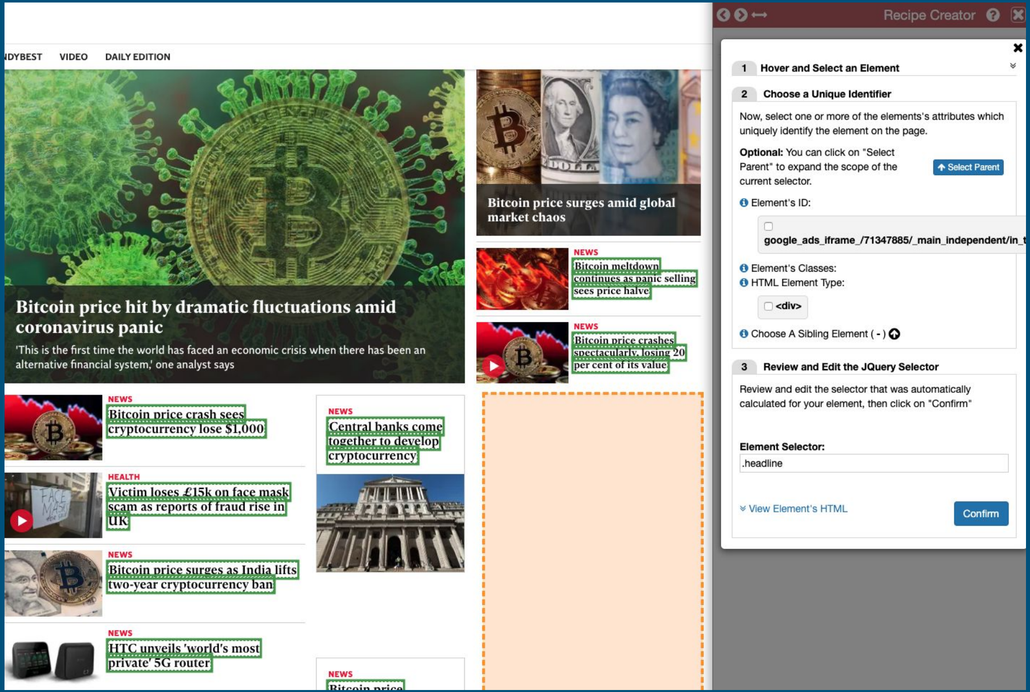
- VWAP- Chainrider Finance API
  - Offers six months worth of data on 16 cryptocurrencies across 10 exchanges
  - Free
  - Fast developer access with quality support
- API Request:

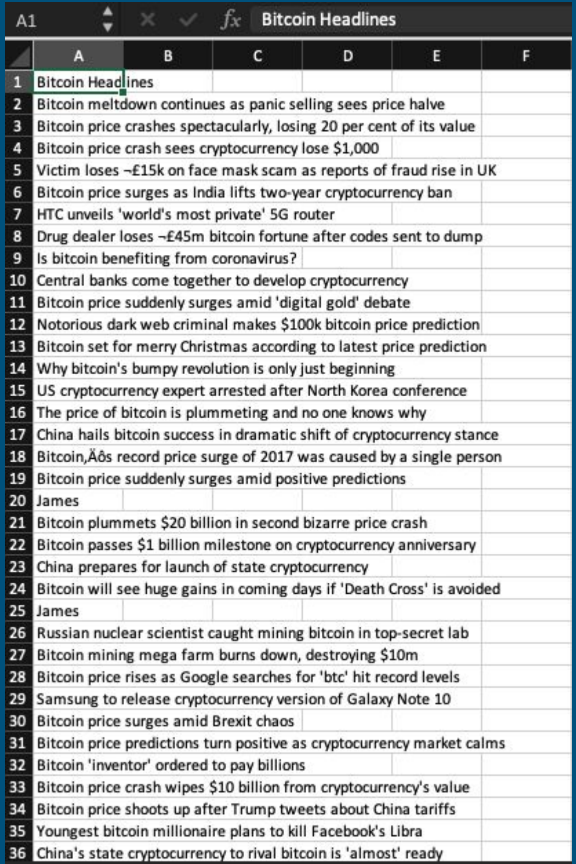
```
body = {
    "pair": "BTCUSD", #listOfCurrencies[i]
    "upper_unix": currentTime,
    "lower_unix": pastTime,
    "analytics": True,
    "exchanges": ["Huobi"] #Exchange we are getting VWAP from
}

#Retrieving VWAP score:
r = requests.post('https://api.chainrider.io/v1/finance/vwap/historic/', json=body,
```

Headline Sentiments

- News Sources: Independent/Coindesk
  - Easy to scrape
  - Popular
  - Articles included timestamps (manual insertion into dataset)
- Data Miner (Web Scraper)
  - Free
  - Automatically places headlines in Excel file
  - 5-10 articles per day from 2/1/20 - 3/8/20

Headline Sentiments





A1 — Bitcoin Headlines

	A
1	Bitcoin Headlines
2	Bitcoin meltdown continues as panic selling sees price halve
3	Bitcoin price crashes spectacularly, losing 20 per cent of its value
4	Bitcoin price crash sees cryptocurrency lose $1,000
5	Victim loses ¬£15k on face mask scam as reports of fraud rise in UK
6	Bitcoin price surges as India lifts two-year cryptocurrency ban
7	HTC unveils 'world's most private' 5G router
8	Drug dealer loses ¬£45m bitcoin fortune after codes sent to dump
9	Is bitcoin benefiting from coronavirus?
10	Central banks come together to develop cryptocurrency
11	Bitcoin price suddenly surges amid 'digital gold' debate
12	Notorious dark web criminal makes $100k bitcoin price prediction
13	Bitcoin set for merry Christmas according to latest price prediction
14	Why bitcoin's bumpy revolution is only just beginning
15	US cryptocurrency expert arrested after North Korea conference
16	The price of bitcoin is plummeting and no one knows why
17	China hails bitcoin success in dramatic shift of cryptocurrency stance
18	Bitcoin‚Äôs record price surge of 2017 was caused by a single person
19	Bitcoin price suddenly surges amid positive predictions
20	James
21	Bitcoin plummets $20 billion in second bizarre price crash
22	Bitcoin passes $1 billion milestone on cryptocurrency anniversary
23	China prepares for launch of state cryptocurrency
24	Bitcoin will see huge gains in coming days if 'Death Cross' is avoided
25	James
26	Russian nuclear scientist caught mining bitcoin in top-secret lab
27	Bitcoin mining mega farm burns down, destroying $10m
28	Bitcoin price rises as Google searches for 'btc' hit record levels
29	Samsung to release cryptocurrency version of Galaxy Note 10
30	Bitcoin price surges amid Brexit chaos
31	Bitcoin price predictions turn positive as cryptocurrency market calms
32	Bitcoin 'inventor' ordered to pay billions
33	Bitcoin price crash wipes $10 billion from cryptocurrency's value
34	Bitcoin price shoots up after Trump tweets about China tariffs
35	Youngest bitcoin millionaire plans to kill Facebook's Libra
36	China's state cryptocurrency to rival bitcoin is 'almost' ready

Headline Sentiments

- Used Textblob library in Python to derive sentiment from headlines
  - Easy implementation
  - Good reputation [4]
- Outputted to Excel file

```python
def analyze_sentiment(self, headline):
    analysis = TextBlob(self.clean_headline(headline))
    if analysis.sentiment.polarity > 0:
        return 1
    elif analysis.sentiment.polarity == 0:
        return 0
    else:
        return -1
```

Bitcoin Headlines	Sentiment
Bitcoin meltdown continues as panic selling sees price halve	-1
Bitcoin price crashes spectacularly, losing 20 per cent of its value	-1
Bitcoin price crash sees cryptocurrency lose $1,000	-1
Victim loses ¬£15k on face mask scam as reports of fraud rise in UK	-1
Bitcoin price surges as India lifts two-year cryptocurrency ban	1
HTC unveils 'world's most private' 5G router	0
Drug dealer loses ¬£45m bitcoin fortune after codes sent to dump	-1
Is bitcoin benefiting from coronavirus?	1
Central banks come together to develop cryptocurrency	0

Tweet Volume & Sentiment

- Tweetbinder
- Received a 5-week supply of data
  - Tweet Volume, sentiment
- Live-streaming tweets myself was challenging
  - Bad requests led to multiple Twitter bans
  - Time
  - Computer would need to run 24/7 (cloud alternatives expensive)

	Total tweets
Neutral	24,158
Positive	9,233
Negative	1,609

$$\frac{\text{(Positive Tweets)} - \text{(Negative Tweets)} + 0}{\text{Total Tweets in Interval}}$$

Final Dataset

- Training set
  - Date Range: 2/1/20 - 3/8/20
  - 266 intervals
- Test set
  - Date Range: 3/27/20-4/1/20
  - 40 intervals (cut to 28 after evaluation)

1	Timestamp	Tweet Volume (Every 3.25 hours)	Average Tweet Sentiment	Average Headline Sentiment	VWAP Score of Interval (3.25 hours)
2	1580576460	114	0.67	0.22	9387.04
3	1580588160	76	-0.18	0.32	9384.03
4	1580599860	104	-0.91	-0.68	9315.75
5	1580611560	119	0.19	-0.77	9285.12
6	1580623260	130	-0.81	-0.81	9356.28
7	1580634960	89	-0.53	-0.95	9423.65
8	1580646660	86	-0.25	0.42	9426.93
9	1580658360	62	-0.85	0.97	9436.94
10	1580670060	65	-0.6	-0.04	9438.97

Step 3- Fitting the Model

- Fed dataset into an LSTM Neural Network in Python
- Used LSTM model in Keras library:
  - **1)** Loaded in dataset
  - **2)** Feature scaling (0-1)
  - **3)** Split data into 90% train, 10% test
  - **4)** Reshaped input data to be 3D (LSTM takes in 3D input)
  - **5)** Built the LSTM model
  - **6)** Transformed outputs back from feature scaling
  - **7)** Visualized results

```python
dataset = pd.read_excel('FINALLSTMDatasetBitcoin.xlsx', nrows = 305) #FINALLSTMDatasetB

values = dataset.iloc[:,1:5].values #Getting vwap scores [2:5]
values = values.astype('float32')

#Feature Scaling- converts vwap scores into values ranging from 0 to 1 (normalizing dat
scaler = MinMaxScaler(feature_range = (0,1))
scaled = scaler.fit_transform(values)

#Retrieve data from previous timestep (Supervised Learning)
reframed = series_to_supervised(scaled, 1, 1)

#Drop columns we don't want to predict
reframed.drop(reframed.columns[[4,5,6]], axis = 1, inplace = True)
print(reframed.head())

#Splitting data into train and test sets
reframedValues = reframed.values
n_train_days = 266 * 1 #90% data is train, 10% test
train = reframedValues[:n_train_days, :]
test = reframedValues[n_train_days:295, :]

#Assigning inputs and output datasets
train_X, train_y = train[:, :-1], train[:, -1]
test_X, test_y = test[:, :-1], test[:, -1]

#Reshaping input to be 3 dimensions (samples, timesteps, features)
train_X = train_X.reshape((train_X.shape[0], 1, train_X.shape[1]))
test_X = test_X.reshape((test_X.shape[0], 1, test_X.shape[1]))
print(train_X.shape, train_y.shape, test_X.shape, test_y.shape)

#Building LSTM Neural Network model
model = Sequential()
model.add(LSTM(50, input_shape=(train_X.shape[1], train_X.shape[2]))) #Recurrent Layer
model.add(Dropout(0.4)) #Dropout Layer
model.add(Dense(20,activation= 'tanh')) #Fully Connected Layer
model.add(Dense(1,activation='sigmoid')) #Output Layer
model.compile(loss='mae', optimizer= 'adam', metrics=['acc']) #Compiling the model

```

```python
#Fitting model
history = model.fit(train_X, train_y, epochs = 180, batch_size=20, validation_data=(t

#Plotting training loss vs validation loss
plt.plot(history.history['loss'], label='train')
plt.plot(history.history['val_loss'], label='validation')
plt.legend()
plt.show()

#Model making a prediction
yhat = model.predict(test_X)
test_X = test_X.reshape((test_X.shape[0], test_X.shape[2]))

#Inverting data back from feature scaling
inv_yhat = concatenate((test_X[:, :-1], yhat), axis=1)
inv_yhat = scaler.inverse_transform(inv_yhat)
inv_yhat = inv_yhat[:,3] #2

test_y = test_y.reshape((len(test_y), 1))
inv_y = concatenate((test_X[:, :-1], test_y), axis=1)
inv_y = scaler.inverse_transform(inv_y)
inv_y = inv_y[:,3] #2

#Calculating RMSE and MAE
rmse = sqrt(mean_squared_error(inv_y, inv_yhat))
mae = mean_absolute_error(inv_y, inv_yhat)
print('Test MAE: %.3f' % mae)
print('Test RMSE: %.3f' % rmse)

#Visualising Results (Actual vs Predicted)
plt.plot(inv_y, color = 'red', label = 'Actual Bitcoin VWAP')
plt.plot(inv_yhat, color = 'blue', label = 'Predicted Bitcoin VWAP') #[1:38]
plt.title('Bitcoin VWAP Prediction')
plt.xlabel('Time Interval (1 interval = 3.5 hours)')
plt.ylabel('VWAP')
plt.legend()
plt.show()

```

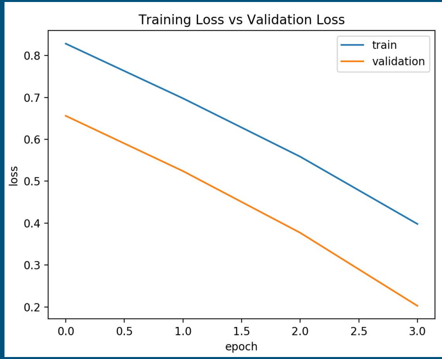
Step 4- Evaluating the Model

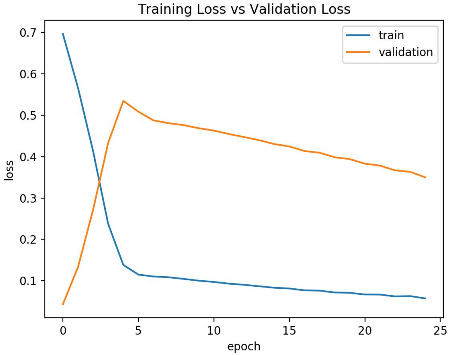
- Root Mean Squared Error (RMSE) of model
  - Standard deviation of the residuals (prediction errors)
  - Lower the score the better
  - Low score can also indicate overfitting- need an additional evaluation method

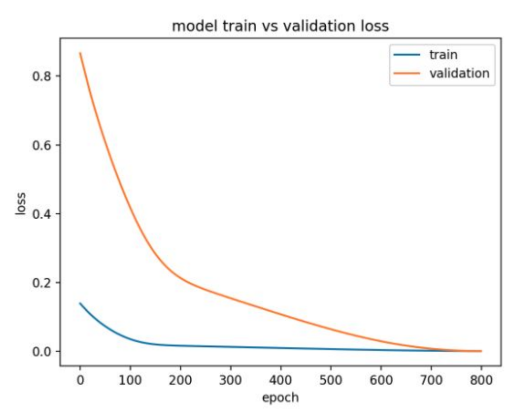
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Step 4- Evaluating the Model

- Graphed Training vs Validation Loss
    - How well algorithm is modeling data
    - Loss- want a low number
    - Underfitting
        - Large gap in between lines
        - Loss doesn't stop decreasing
    - Overfitting
        - Lines converge then seperate
        - Validation loss is volatile (doesn't stabilize)
    - Ideal Model
        - Two lines trend downward, eventually converging and stabilizing at a fixed loss
        - Points do not intersect until end of graph

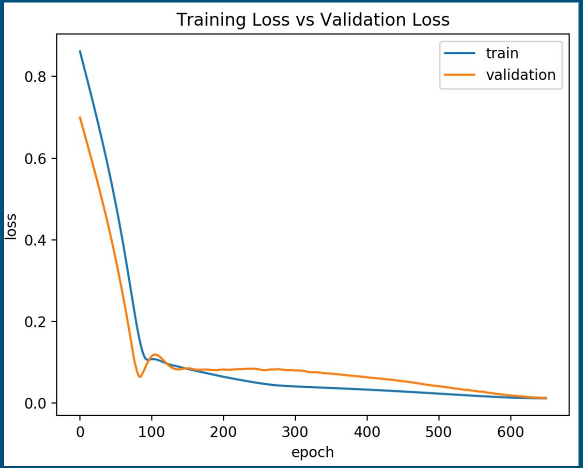


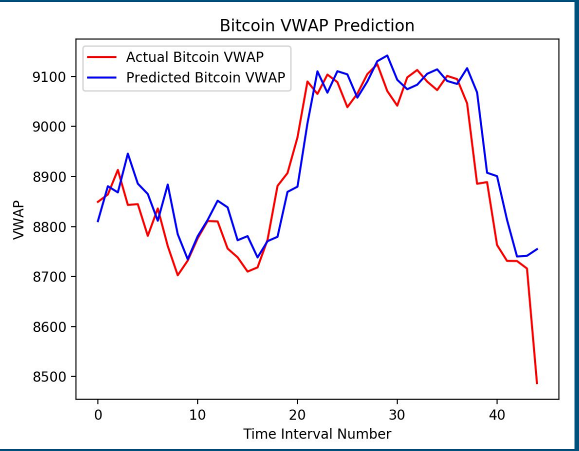




Step 4- Evaluating the Model

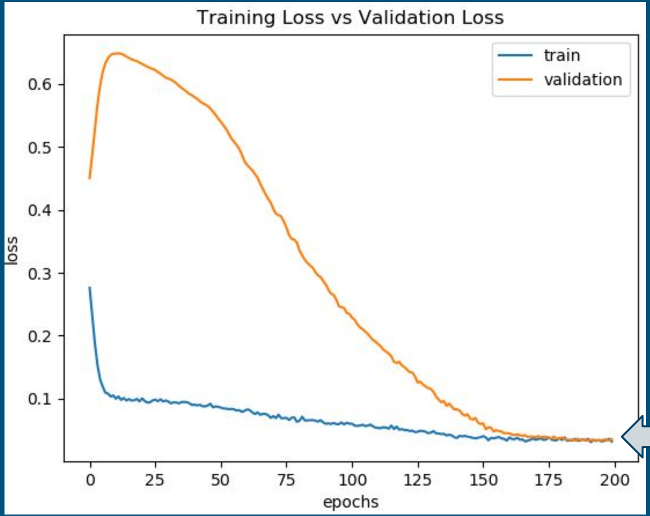
- Initial Problems: Overfitting, Underfitting, high RMSE
  - Decreased test set to 10%
  - Changed activation function (hidden layer used tanh, output layer used sigmoid)
  - Increased dropout rate to 0.4
  - Adjusting number of epochs, batch size, neurons in each layer- trial and error
- Goal: Good fit on graph while maintaining the lowest RMSE score possible





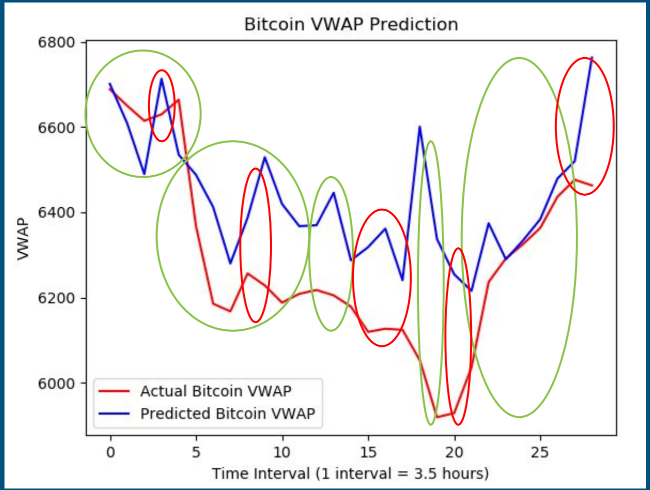
Results

- Best results:
  - Epochs- 200
  - Batch Size- 32
  - Dropout rate- 0.4
  - Recurrent layer- 50 neurons
  - FCL- 15 neurons, tanh activation function
  - Output Layer- sigmoid activation function
- Lowest RMSE- 160.04
  - Model fits well, RMSE can be lowered with a larger sample size
- Model predicted 22% more of the intervals correctly with sentiment data vs. model without



Loss decreases steadily

Lines converge and stabilize



Predicted 21/28 intervals correctly (75.00%)

Future Work

- Will work on live streaming tweets to increase dataset
    - Higher accuracy
- Make LSTM models for other cryptocurrencies like Ethereum, Litecoin
    - Try to find patterns between cryptocurrencies
    - Do some cryptos influence the market of others?

Thank You!

- Dragan Boscovic- Thesis Director
- Hasan Davulcu- Thesis Committee Member
- John Billings- Barrett Honors Advisor

References

[1] Brownlee, Jason. “How to Diagnose Overfitting and Underfitting of LSTM Models.” *Machine Learning Mastery*, 7 Jan. 2020, machinelearningmastery.com/diagnose-overfitting-underfitting-lstm-models/.
[2] Kang, Eugine. “Long Short-Term Memory (LSTM): Concept.” *Medium*, Medium, 1 Sept. 2017, medium.com/@kangeugine/long-short-term-memory-lstm-concept-cb3283934359.
[3] Shearer, Elisa, and Katerina Eva Matsa. “News Use Across Social Media Platforms 2018.” *Pew Research Center's Journalism Project*, 31 Dec. 2019, www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/.
[4] Pant, Neelabh. “A Guide For Time Series Prediction Using Recurrent Neural Networks (LSTMs).” *Medium*, Stats and Bots, 7 Mar. 2019, blog.statsbot.co/time-series-prediction-using-recurrent-neural-networks-lstms-807fa6ca7f.
[5] Jain, Shubham. “Natural Language Processing for Beginners: Using TextBlob.” *Analytics Vidhya*, 5 Sept. 2019, www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/.
[6] Mitchell, Cory. “Volume Weighted Average Price (VWAP) Definition.” *Investopedia*, Investopedia, 2 Mar. 2020, www.investopedia.com/terms/v/vwap.asp.
[7] Brownlee, Jason. “Difference Between a Batch and an Epoch in a Neural Network.” *Machine Learning Mastery*, 25 Oct. 2019, machinelearningmastery.com/difference-between-a-batch-and-an-epoch/.
[8] Zhang, Lena. “Why Investing in Bitcoin Is Hot Again.” *AllTechAsia*, 19 Sept. 2019, alltechasia.com/why-investing-in-bitcoin-is-hot-again/.
[9] Gradojevic, Nikola. “The Answer to Forecasting Bitcoin May Lie in Artificial Intelligence.” *The Conversation*, 11 Dec. 2019, theconversation.com/the-answer-to-forecasting-bitcoin-may-lie-in-artificial-intelligence-119152.
[10] Galactic, Virgin, and Chamath Palihapitiya. “Bitcoin User Demographics: European Males Age 25-34: News Bitcoin News.” *Bitcoin News*, 19 Sept. 2016, news.bitcoin.com/bitcoin-user-demographics-european-males-age-25-34/.