Predicting Prices of Oil Using Twitter Sentiment and Long Short Term Memory

Andy Vu

Abstract—Since its fruition with the onset of the industrial revolution, the price of oil per barrel has fluctuated greatly as it found its way to become the number one commodity. Throughout the years, the price of oil has been difficult to predict among investors and industry experts alike. Other than the simple economic principle of supply and demand, there are other factors that have shown to be influential to oil prices. These factors such as war, global pandemics, depleting reserves, and environmental concerns have pushed and impacted the price of oil in some manner. The aim of this project is to propose a novel approach to predict the price of oil. Using social media platforms as a way to organize and collect the thoughts of the masses and to determine the outlook on oil, our proposed method identifies the trend of oil prices using these sentiments. Social media giants such as twitter offer a vast sea of untapped sentiment data. These tweets can be interpreted for their sentiment by using natural language processing. Combining this sentiment data and when paired with a long short term memory recursive neural network, there are promising results to show a strong correlation with WTI oil prices and sentiment data. With our approach we were able to obtain an R-squared value of 0.831121 with our inputs and WTI crude oil price.

Index Terms—Machine Learning, Oil, Twitter, Sentiment, Data Analysis, Time series, Price Prediction, Stocks, Commodities, Social Media

1 Introduction

Oil is often called black gold or liquid gold, due to its inherent value as well as its impact on global society. At the dawn of the industrial revolution as humans were starting the mass use of fossil fuels, a barrel of oil would fetch a price of roughly US$1.19 [1] As the usage of these fossil fuels spread from America to the rest of the world, the price increase of barrel of oil continues to rise and drop. At its peak, oil reached a record high of $150 per barrel moments before the 2008 global economic recession. On the opposite hand to this, during the global Covid-19 pandemic, for the first time ever in history, prices of a futures market went negative. Oil futures prices plummeted into negative prices of -$40 for two days. Today's global worth of oil and its market are estimated at a seven and a half trillion dollars [2].

It is without a doubt that oil has become a critical commodity for many counties today with some counties entirely having their entire economy dependent on the natural resource. Countries such as Saudi Arabia rely on oil so heavily that 40% of the country's GDP is from the oil sector with oil accounting for roughly 85% of the country's exports [3]. Another example is the Country of Norway. Norway's export of oil and gas accounts for 45% of the country's export while those products contribute to 20% of the GDP [4]. Another example to help emphasize the importance of global oil is that the 30% of the entire European Union receives oil imports from Russia [5]. These countries are just a few of the many throughout the world that has large oil and natural gas sectors.

The study of the price of oil and its historical price is important not only to investors, economist and industry experts. Because of petroleum products that trickles its way from starting its life as crude oil to refined products, every human on this planet uses oil products or byproducts [6]. Products such as gasoline at the gas station fluctuates its price on supply and demand but also its price is directly affected by per barrel price of oil. Also because the price of oil per barrel directly affects the everyday citizen, it has been found that every time oil exceeds its high by 100% a recession occurs [7]. At the time of this writing, due to the Russian-Ukrainian conflict, oil has found itself to increase its prices over 100% at a peak high of $130 in March from $65 in December. The coming of economic recession and analysis of oil prices is critical to the common man as much as it is important to economist and financial advisers.

With the concern of global oil prices, it is crucial that we are able to have estimations or predictions in regards to the trends of the oil markets as any other market. There has been countless indicators, formulas and approaches to predict the price of commodities and assets ever since trading has occurred in human history. In today's time experts use technical analysis combined with historical data to also predict future prices with varying success [8]. In addition to this, market experts combine technical analysis with fundamental analysis to predict future prices as well with similar results [9]. Many of these tools and strategies can be found in the late 20th century and are applied to the equities market [8].

This paper explores a relatively novel approach to predict oil prices by using machine learning. Natural language processing, and social media. Fundamental news catalyst that drive market trends or push price in a certain direction are often tweeted about. Financial news channels and individuals post information regarding critical economic deci-

A. Vu is with the College of Computing and Software Engineering, Kennesaw State University, Marietta, GA, 30060 USA. Emails: avu5@students.kennesaw.edu

sions such as interest rate hikes or monthly unemployment data. On top of this there are many industry experts and investors that vocalize their opinions on the twitter social media platform giving insights and elaborations upon these financial news statements. By collecting these tweets and analyzing the contents for sentiment, we can combine this data with machine learning techniques to perform price prediction.

2 Related works

This approach to using twitter sentiment or other social medial platform sentiment to predict future prices of assets is not entirely new. Although in comparison to other conventional methods that have been used to prior to social media it is relatively new. Twitter was launched in 2006 however there were not many users until later. Because of this, only until recently was this endeavor feasible. One of the most early attempts was by Bollen [10]. Here Bollen and team used twitter moods categorizing the sentiment into multiple categories other than positive, negative or natural. Here they used calm, alert, sure, vital, kind, and happy [10]. Another interesting note is that Bollen used twitter sentiment data against the Dow Jones Industrial. As a culmination of the 30 leading industry leading company stocks it allows a more averaged view compared to the sentiment against a single particular stock or entity. It could be said that using the Dow Jones Industrial as an example broadens the scope as it is a strong representation of the American equities market and in a sense the performance metric of the American economy.

In a similar approach, Sattarov, uses twitter sentiment to forecast bitcoin price fluctuations. However, their sentiment analysis is only of a data set of 92,550 tweets, in a span of 60 days [11]. In addition to this, their forecast is not a precise price point but rather a range of prices to show a correlation. Because the price fluctuations are a much broader range it could possibly be more feasible to predict a correlation. Another aspect of their design is the use of VADER (valence aware dictionary and sentiment reasoner) which may be a more powerful social media sentiment analysis tool compared to the textblob analysis used in our paper which is more discuss below in the approach section. Despite the difference in the text analysis, both output a score from -1 to 1 in polarity [11] [12] [13].

Oussalah and Zaidi were able to successfully show the correlation between WTI oil prices with the use of twitter sentiment focusing on United States foreign policy tweets as well as oil company sentiment [14]. In contrast to the aim of this paper which is more broad, Oussalah and Zaidi had a more focused twitter analysis [14]. They made use of other sentiment analysis tools that includes SentiStrength, and Stanford NLP sentiment. With this, they used a variety of machine learning techniques but found support vector machines to have the best results after a forecasted price direction of seven days [14].

Like Oussalah and Zaidi, Sadik et. al focused their news sentiment analysis in a narrow field. Their macroeconomics used the RavenPack News Analytics which delivers sentiment analysis and news event data which are most likely to impact financial markets and trading [15]. With these services, Sadik and team focused only on news related to crude oil price movements which include companies, organizations, currency, commodity and people [15]. In addition to this and other restrictions, they limited the events to only the top 15 countries with oil imports and exports [15]. In a comparison of using one-factor model with and without the news data they were able to obtain a RSME score of 2.67 and 3.15 respectively [15].

To compare between sentiment data between multiple sources, Elshendy et. al focused on analysing oil prices using four different social media networks [16]. They used a variety of sources for sentiment and news related data. For twitter, they used tweets, tweet sentiment, tweet emotionality, and tweet complexity. As for Google trends, they used ”OPEC” query count, as well as the ”Price of Oil” query count. Wikipedia ”OPEC” page views and ”price of Oil” page views were considered as well. Finally they used GDELT number of organizations and number of articles [16]. By comparing these different platforms, Elshendy et. al found that twitter was the most predictive given a one day time lag while google trends and Wikipedia sources required more time and are most effective after a 3 day time period. It seems as though twitter has a higher potential for traders and speculators to voice their opinions.

This paper, takes direct references and inspiration from Pagolu et. al. who also took inspiration from Bollen [17]. One of the differences is the choice of market instrument. In this case, Pagolu uses twitter data to predict stock movements. Although similar, oil is a much larger commodity and is much more linked to fundamental movements as it is more globally interested compared to Apple as a single stock in the New York stock exchange. Using code and github from you915 [18], this project achieves a similar outcome. You915 was able to use a data set of Apple stock to predict the future price movements with some success which is further discussed in the results and discussion section. There are also other projects which have become the inspiration for this paper such as twitter sentiment analysis for bitcoin, and other stocks are which are often done. Despite this, there seemed to be fewer papers and readily available datasets that use twitter sentiment to predict the price of commodities such as gold or oil whereas most machine learning based projects tend to gravitate towards stocks, cryptocurrencies or industrial indexes such as the Dow Jones or S&P 500.

Most of the related works often used LSTM RNN models to the highest accuracy compared to the other models that were used such as regression or decision trees. As can be seen by you915, in addition to a LSTM, they used a random forest regression where the results are not ideal [18]. The predicted values of the Apple stock are at some points close to the actual value while most are extremely off. Utilization of this random forest regression tree allows us to see that some models are clearly more superior than others. This is not to say that other models which are not LSTM RNN are not effective. Li, et. al gathered news sentiment using a convolution neural network (CNN) and was able to use a support vector regression to forecast oil prices which proved to be more effective than their random forest or linear regression [19]. An example of different model strengths is the work done by Wu et. al [20]. Here Wu et. al predicted oil prices during the global pandemic using multiple models:

backpropagation neural network, multiple linear regression, support vector machine, LSTM and RNN [20]. Of these models, depending on which features were selected showed that each model has their own strengths and weaknesses [20]. For these reason, we have chosen to forgo even a baseline model of regression or other simple methods and instead went directly with a LSTM RNN approach.

Other interesting models include the hybrid model proposed by Zhenda Hu [21]. Here, Hu combined a complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) model with a LSTM RNN. The EMD is used for nonlinear and nonstationary signals but often times results in mode mixing problems [21]. To combat this, the CEEMDAN is used to more accruately obtain mode functions [21]. Hu combined the CEEMDAN signal processing with the LSTM to use for time series predictions and was able to obtain results that bested a standalone LSTM model as well as a random forest model [21].

3 Approach (and technical correctness)

On average, there are roughly 500 million tweets per day as of 2016 growing from 50 million since 2010 [22]. As of 2019 twitter had roughly 330 million monthly active users growing ten times as much since 2010 [23]. It is easily understood that as time goes on, twitter will accumulate more users and thus more tweets will be tweeted daily. With this in mind, twitter is an open sea vast with opinions and facts both useless and vital as well as constantly growing. Using these tweets, specifically ones that contain valuable data, to create a dataset that is analyzed for sentiment allows us to create an informed prediction for oil prices. Both Pagolu, you915 and others often use pre-established data sets that already includes stock prices with their respective detailed information such as open price, close price, adjusted volume, as well as twitter polarity and twitter volume.

These data sets are often available for stocks but we were not able to find an appropriate data set for oil. Also there are also other works that generate their own twitter data sets and merge this data with the correlating commodity of interest. Due to this paper focusing on oil futures market instead of the equities market and lack of already available data sets, we are thus forced to create an appropriate data set related to oil.

Creating a twitter scraper function is relatively straight forward and simple using the tweepy python library. Firstly, a twitter developer account is needed which provides access keys to interface with twitter and our built application programming interface (API). Despite this however, there are several limitations regarding the usage of twitter's API. Since the middle update of this project, we have been able to overcome a key obstacle of obtaining an academic research twitter developer account. This account research privileges allows the application to search all tweets dating back to the very first tweet since 2006, while in comparison, elevated or basic access privileges only allow searches of tweets less than seven days old [24]. In addition to this, academic research privileges allow a tweet capacity of 10 million every month in comparison to two million [24]. Because of the nature of this time series analysis, there is a requirement of tweets being older than seven days. In addition to tweet capacity limitations, there are rate limits imposed on the application as well. Using the full archive search function, applications are limited to 300 request per every 15-minute window with an additional rate limit of one request every second [25]. Each request is limited to a maximum of 500 results for every search query [24].

Even with academic research access privileges the scope of the project far exceeds the limitations of any free twitter access. Pulling every tweet since 2010 with the keyword oil would far exceed the tweet capacity even if the rate limitations are disregarded. This would result in over 4483 days. Assuming there are 5000 daily tweets regarding oil, this would be approximately 22 million tweets. Because of this, the rate limitations, and max query results, we have decided to specify our query to be oil, gas, and wti(west texas instruments) all connected by the 'and' search operator. We also specify that no retweets are pulled. By searching in this order of specificity, this allows each search query to be meaningful and unique. The downside to this is that, by excluding retweets we are limiting the volume of the tweets. There could be thousands of retweets for a single unique tweet that would show the significance of that tweet. Also by including wti and gas, it ensures that the searches pulls up results of the oil commodity and not oil products such as sunflower oil, cooking oils, or automobile oils. Our tweet function, searches for tweets since 2010, with the above specified search queries pulling 500 tweets, waiting a second between each request and then also waits 15 minutes every 300 request. This results in only 500 tweets on a specified search day. This process is iterated using the python datetime function from December 2nd 2009 until March 2nd 2020 [26].

Once these tweets are pulled and written into a csv (comma separated values file, they are then to be processed to be cleaned using regular expressions [27]. The tweets are cleaned removing @mentions, the # symbol, RT(retweets), and hyperlinks. From here, the polarity of the tweets are analyzed using TextBlob. TextBlob is a python natural language processing library that has a package that can perform sentiment analysis resulting in the polarity and subjectivity of the input text [12]. For our application, subjectivity is not of any concern and thus we only use the TextBlob sentiment analysis for polarity. Polarity is given a score ranging from -1 to 1 with -1 defining negative sentiment while a 1 defines a positive sentiment [13]. Further analysis of the tweets are to count the amount of negative tweets and positive tweets, while tweets that are neutral show a polarity of 0. In cases where there are no tweets that result from a query for a specific day, such as in the early times of twitter in 2010, with low amounts of tweets and users, the null values are also represented as zeroes and thus neutral polarity. From here, the tweets are then averaged by day resulting in a daily average polarity and then merged by date to match with the wti commodity that can be found on investing.com [28]. In addition to the polarity, the twitter volume, is how many tweets were found for a given day is also recorded with the corresponding date. Lastly, non-trading days are dropped from the dataframe resulting in a final csv file containing polarity, twitter volume, and oil price sorted by date. All of the functions and code used to pull tweets, as well as process the tweets into the final

CSV for output that is fed into the LSTM model is all made from scratch and original aside from using packages such as tweepy, TextBlob, and Pandas.

Once the final csv file is obtained with daily close prices of oil matched with the daily sentiment polarity, a variety of machine learning models can be used to interpret the data and to forecast the future prices of oil. For the purpose of this paper, we have gone with a long short term memory (LSTM) recurring neural network (RNN). While the reference project of you915 chose both an LSTM as well as a random forest regressor and other projects often use some type of regression as well [18]. The LSTM model of our choice uses tensorflow libraries for ease of use [29]. The features used are the close price, the polarity, and the twitter volume. As for training and testing purposes the model is split 70% training while the remaining 30% is for testing. This leaves 638,269 tweets for training and 273,544 tweets for testing for the first dataset, while for the second dataset with this same percentage of split results in 1,528,556 tweets for training and 655,096 tweets for testing. In addition to this the training and testing data are also scaled using the module MinMaxScaler from sklearn [30].

As for details of the LSTM model, which uses some baseline and then modified parameters from you915, it uses three hidden layers with a final output layer which uses an Adam optimizer while focusing the loss function as mean squared error. This three layer LSTM model allows a three day window of previous closing prices to be used to estimate the future price. For our purposes, we have increased the window day period to five days instead of three to further allow a full trading week's worth of data to be used instead of the previous three days. Each of the LSTM layers have nine neurons within their respective layers and use a dropout fraction value of 0.2. The dropout function is included to assist with any over fitting that may occur by probabilistically excluding activation and weight updates in the network. The model is trained with batch sizes of five along 10 epochs. For a second run, we increase training to 100 epochs. An example summary of the model is shown in table 1 below.

TABLE 1: Model summary of the LSTM RNN used to forecast the next three days prices.

Model Summary		
Layer (type)	Output Shape	Param #
lstm (LSTM)	(none, 9, 9)	396
dropout (Dropout)	(none, 9, 9)	0
lstm_1 (LSTM)	(none, 9, 9)	684
dropout_1 (Dropout)	(none, 9, 9)	0
lstm_2 (LSTM)	(none, 9)	684
dropout_2 (Dropout)	(none, 9)	10
Dense (Dense)	(none, 1)	0

4 Experimental results (and technical correctness)

Using the search queries of 'oil gas wti' and excluding retweets, the tweet search resulted in 911,813 total tweets ranging from dates of December 2nd 2009 to march 2nd 2022. This is half of the expected amount of tweets that were calculated in the approach section. 4450 days multiplied by 500 daily tweets is an expected 2,225,000 tweets. Our resulting dataset of half the expected tweets can be seen due to the earlier days of twitter history. There are many pulled tweets where for many dates, there are zero tweets that resulted from the search. This could be attributed to the popularity and growth of twitter. Days much more closely to the present time have a much higher volume count compared to days closer to 2010. This can be seen in table 2. The volume of tweets is much lower a decade ago compared to today. Cleaning the tweets and using the TextBlob sentiment analysis resulted in a surprisingly amount of negative tweets. Of the total tweets, 561,315 were found to be scored negative, with 154,282 tweets given a polarity score of zero either from no found tweets from that day, or from being a true neutral tweet. Lastly, there were 196,215 tweets that were found to be positive. An example of tweets, with their polarity can be seen in table 3.

TABLE 2: Example data of final dataframe created from a csv file to be fed into the LSTM model after tweet cleaning and average polarity is calculated. This uses the first tweet dataset with less than 1 million tweets.

Final Oil CSV			
Date	Polarity	Volume	Price
2009 − 12 − 02	−0.43333	1	76.60
2009 − 12 − 03	0.03333	4	76.46
2009 − 12 − 04	−0.43333	3	75.47
....	....	.....	.....
2022 − 02 − 28	−0.15810	272	95.72
2022 − 03 − 01	−0.05396	329	103.41
2022 − 03 − 02	−0.06313	258	111.38

TABLE 3: Example of shortened tweets with their respective dates and polarity shown. These tweets were shorted to the first few words. This uses the first tweet dataset with less than 1 million tweets.

Tweet Polarity		
Date	Text	Polarity
2009 − 12 − 02	″In the energy...″	−0.43333
2009 − 12 − 31	″Oil & Gas - Mark...″	0
2010 − 01 − 06	″New blog post...″	−0.28181
....	″................″	....
2020 − 04 − 20	″Great Job Mr Pr...″	0.13
2022 − 02 − 18	″Russian Foreign...″	−0.03214
2022 − 02 − 24	″WTI just touch...″	0.233

Seeing as there was a lower than expected count of tweets, we decided to re-pull tweets using the search query of oil only. This time we also emphasized that tweets were to not be retweeted. The tweet search resulted in almost every search day resulting in nearly the 500 max result. With a total of 2,183,652 tweets which is over double the amount of the first tweet search. Further analysis of these tweets show that there was 369,426 tweets that had a negative polarity, while 989,729 tweets had a polarity score that was neutral leaving a total of 824,497 tweets with a positive polarity. This a much different change in comparison to the first data set. However, due to the scale of the tweets we are not able to verify if all tweets are truly talking about the commodity oil and other oils. Table 4 shows a comparison between the two datasets that were obtained.

TABLE 4: Data comparison between the two datasets obtained for oil using different queries.

Dataset Comparison		
Category	Oil Dataset 1	Oil Dataset 2
Dates	2009-12-02 - 2022-03-02	2009-12-31 - 2022-03-02
Total tweets	911, 813	2, 183, 652
Mean twitter volume	258.566	490.787
Negative tweets	561, 315	369, 426
Negative days	2341	25
Neutral tweets	154, 282	989, 729
Neutral days	605	0
Positive tweets	196, 215	824, 498
Positive days	238	3139

Using baseline features for the LSTM RNN model that can be found in you915's model for analyzing apple stock we were able to output a prediction for future oil price after a 70% training 30% testing split using the first tweet dataset. A defining feature was the training of five batches in 10 epochs. We were able to obtain a root mean square error (RSME) of 0.23351 alongside an R-squared error of -4.09788. These two error values show that there is significant room to improve the model especially with a negative R-squared error value. Below is table 5 showing several values of real prices compared to predicted oil prices. In addition to this, there is figure 1 to show the real prices of oil in comparison to the predicted values plotted on a chart.

TABLE 5: Predicted prices in comparison to the real prices of WTI oil on a given trading day using 10 epochs of training on the first oil dataset with 1 million tweets.

WTI Oil Predicted vs. Real Prices (10 epochs, dataset 1)		
Date	Predicted	Real
2018 − 07 − 20	26.27	70.46
2018 − 07 − 23	26.07	67.89
2018 − 07 − 24	25.81	68.52
....	.....	.....
2022 − 02 − 28	30.72	95.72
2022 − 03 − 01	30.46	103.41
2022 − 03 − 02	31.02	111.38

Real vs Predicted Price of WTI Oil (10 epochs, dataset 1)

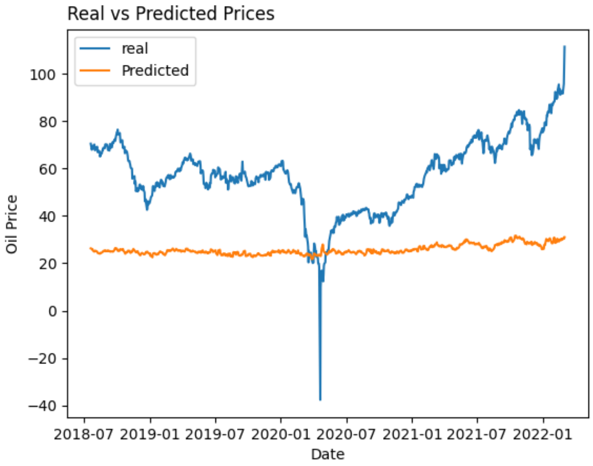

Fig. 1: Plotted chart of the real price of WTI oil in comparison to the predicted prices using 10 epochs of training on the first oil dataset with 1 million tweets.

In comparison with these results, you915 was able to obtain much better results with these standard baseline settings for the LSTM model using the Apple Stock dataset. They were able to obtain a RSME of 0.08039 along side a R-squared value of 0.8842. In addition to this, their predicted vs real prices are surprisingly close for only 10 epochs of training. These real vs predicted values can be seen in table 6 and the baseline model fitting for the Apple stock price can also be more closely observed from plotting these points seen in figure 2.

Real vs Predicted Price of Apple Stock (10 epochs)

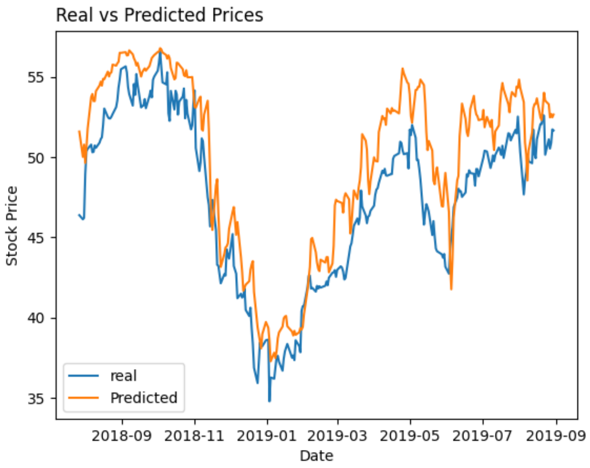

Fig. 2: Plotted chart of the real price of Apple Stock in comparison to the predicted prices using 10 epochs of training

TABLE 6: Predicted prices in comparison to the real prices of Apple stock on a given trading day using 10 epochs of training.

Apple Stock Predicted vs. Real Prices (10 epochs)		
Date	Predicted	Real
2018 − 07 − 30	49.94	46.13
2018 − 07 − 31	48.94	46.22
2018 − 08 − 01	47.64	48.95
....	.....	.....
2019 − 08 − 28	50.35	50.88
2019 − 08 − 29	49.70	51.74
2019 − 08 − 30	49.72	51.67

It was apparent that as can be seen in the table and figures, the predicted prices for WTI oil using the first dataset, are extremely inaccurate not showing a proper prediction. Increasing the training epochs to 100 allowed a great increase in accuracy of the model. In comparison, we were able to obtain a RSME of 0.05053 and an R-squared value of 0.76131. Looking at the figures it is also apparent that the predicted prices are much closer to the actual prices of oil when increasing the training from 10 to 100 epochs.

The results of running the second dataset through the LSTM with baseline settings shows that the difference in data does not make a large impact in terms of accuracy for the model especially with 10 epochs of training. With the same training as the baseline from Apple stock, a poor score of 0.2258 for RSME was acquired alongside a R-sqaured value of -3.7542. The results of the real vs predicted prices which can be seen in their respective figures also show a similar result to outcome of the first dataset. For this reason inclusion of a table to show real prices compared to predicted prices is not included for the second dataset at 10 epochs of training. However, the charted plot of the price points are included in figure 3. After 100 epochs of training, the second dataset obtained scores that were reduced to 0.04255 for the RSME, and 0.83124 for R-Sqaured which can be compared to the scores for the first dataset in table 8.

Real vs Predicted Price of WTI Oil (10 epochs, dataset 2)

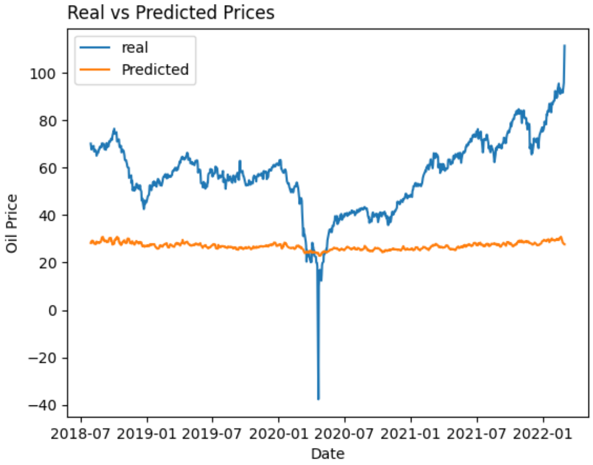

Fig. 3: Plotted chart of the real price of WTI oil in comparison to the predicted prices using 10 epochs of training on the second oil dataset with 2 million tweets.

Real vs Predicted Price of WTI Oil (100 epochs, dataset 1)

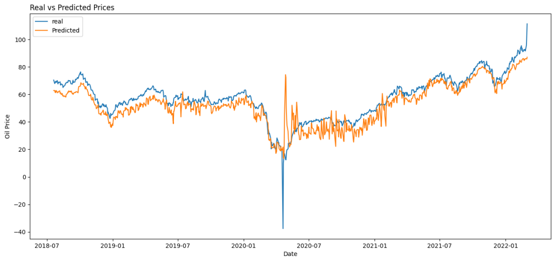

Fig. 4: Plotted chart of the real price of WTI oil in comparison to the predicted prices using 100 epochs of training on dataset 1.

TABLE 7: Predicted prices in comparison to the real prices of WTI oil on a given trading day using 100 epochs of training on the first oil dataset with 1 million tweets.

WTI Oil Predicted vs. Real Prices (100 epochs, dataset 1)		
Date	Predicted	Real
$2018-07-20$	62.81	70.46
$2018-07-23$	62.87	67.89
$2018-07-24$	61.50	68.52
....	.....	.....
$2022-02-28$	85.97	95.72
$2022-03-01$	85.74	103.41
$2022-03-02$	87.10	111.38

TABLE 8: Calculated Loss from the models showing the RSME and R-Squared values obtained after 10 epochs and 100 epochs for baseline Apple Stock and WTI oil for both datasets

Calculated Losses of Apple Stock		
Epochs	Root Mean Squared Error (RSME)	R-Squared
10 epochs	0.1334	.6801
100 epochs	0.09503	.83817
Calculated Losses of WTI oil on Dataset 1		
Epochs	Root Mean Squared Error (RSME)	R-Squared
10 epochs	0.23351	−4.09788
100 epochs	0.05053	0.76131
Calculated Losses of WTI oil on Dataset 2		
Epochs	Root Mean Squared Error (RSME)	R-Squared
10 epochs	0.2258	−3.7542
100 epochs	0.04254	0.831121

Real vs Predicted Price of WTI Oil (100 epochs, dataset 2)

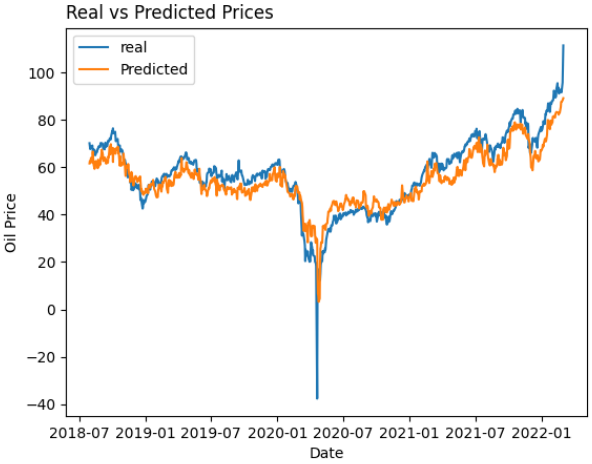

Fig. 5: Plotted chart of the real price of WTI oil in comparison to the predicted prices using 100 epochs of training on the second oil dataset with 2 million tweets.

5 Discussion and Future Works

Looking at the results and data that is shown, it is clear that using twitter sentiment predictions can obtain results to predict the price of a future asset to a degree which in this case is WTI Oil. Of course like any model and test result there is always room for improvement. Our losses calculated from root mean squared error as well as R-squared values can always be reduced. As can be seen by increasing the training epochs, we were able to greatly increase the scores. Modifying other parameters of the model such as increasing the amount of nodes did not create such an impact. On top of this, modifications to the dataset will of course show potential results to obtain better values and reduce error. By creating a larger dataset with larger amounts of tweets and longer historical price data, we could potentially obtain better results. The collected approximate two million tweets is still rather small compared to the large amount of tweets about oil given the daily max of 500. If we were able to obtain every tweet regarding oil and also including retweets this would allow the model work more properly as intended. However, this would require months of data collection as well as other methods to surpass a 500 max result per request. In addition to this, as stated in our results section, more than half of the obtained tweets in dataset one resulted in negative polarization scores. This leads to the question of whether there is an error or bias in the TextBlob packages to score these tweets or is it due to the nature of the tweets and oil to naturally lean towards a negative sentiment. The second dataset had a much different variation of polarity scores which was almost opposite of dataset one with more positive sentiment than negative. In comparison to the baseline apple stock dataset a high majority of the stock polarization scores were positive. Lastly another option would be to forgo the LSTM RNN model altogether and use other models such as regressions or support vector machines for example.

Despite the results however, there is an interesting location where the predicted prices between the first and second dataset are extremely off. This area of interest would be where oil fell to a negative value. It could be seen that this area and for several months after, the model with the first dataset had difficulty to predict the prices of oil. It would seem that the negative and drastic price changes managed to cause the model some great difficulty especially the day after the negative spike. The Predicted price is a large spike in the opposite direction and the days following result in a extremly choppy pattern to predict future prices. In the case for the second dataset, the model was able to more closely follow the price of oil even with the large negative price drop. The days following the price drop, are more steady and closer to the real prices as well with less fluctuations. On top of this, the RSME as well as R-squared values for the second dataset both have higher scores when compared to the first dataset by almost 10% increase. The reasoning behind this is still yet to be discovered but there is speculation that having more positive sentiment tweets is the cause.

In addition to using our own custom dataset, an endeavor we would like to approach is to individually weigh tweets. The rationale behind this is that certain tweets and users contain much more inherent value when it comes to data and information. An example of this is a novice investor that is extremely vocal about their own trading. This individual may post multiple tweets updating their positions or perhaps the individual is uneducated in the topics of finance and economics and post tweets that are incorrect. On the opposite hand, there are many trading funds, and corporations that are extremely well versed in to the economic topic at hand. A tweet from a financial fund that paraphrases the decision of an OPEC (organization of the petroleum exporting countries) chairman could possibly be an immediate sway in the markets that affects central banks and other powerful institutions where a novice trader could not. Without this clear bias and distinction introduced, a novice retail trader would have as much influence on our trading model as an official bank or a president of a corporation. Both tweets from these separate entities would count as sentiment value in the grand scheme of averages.

An example of this logic as follows: the retail trader says oil markets are optimistic to gain value while the OPEC chairman tweets that supply of oil is too high with little demand. The individual trader is proposing to our model as one vote of positive sentiment to indicate a raise in price, while the chairman is an opposite one vote of negative sentiment to lower price. In reality, there are several factors at play. One is that the chairman's tweet could be taken as an actual fact and thus the prices of oil will in fact drop due to the lack of demand and high of supply while the individual trader is simply stating their analysis which may be biased from their positions on oil contracts. Secondly, the reach of the chairman is much wider in audience than the single retail trader. The chairman could have hundreds of thousands of followers and retweets while the single trader only has several followers. As it stands, our current model weights both of these tweets equally out of the many that are tweeted daily. One negative and one positive sentiment count out of a total of 2000 for example. This is completely prone to error as the single one tweet from the chairman could potentially outweigh the other 1998 tweets for the day thus changing the potential polarity. Developing an advanced sentiment system to distinguish these tweets will not be an easy task however it will prove to be extremely valuable in a future model to predict the price.

6 Conclusion

Due to modern inventions and the rise of social media we are presented a new commodity that was not yet available to the common man even twenty years ago: data. Data of itself is useless unless it can be used in a meaningful way. With this said, using twitter as a social media platform and obtaining the data there we are able to predict the price of assets, equities and commodities in ways that was not previously possible several decades ago. This endeavor allowed us to analyze twitter sentiment using twitter API gathering a sum of tweets. These tweets then have their sentiment data obtained using natural language processing. Pairing the obtained twitter data with historical oil prices, we use long term short term memory to successfully predict oil commodity prices to a degree of accuracy. This paper has obtained a R-squared value of .831121 along side a root mean squared error of 0.04252 at best.

Despite the performance not being perfect, the machine learning model is able to be refined to more optimally obtain better predicted future prices. The model is limited in twitter data and also lacks twitter sentiment focus where this of itself could lead to a variety of complications ranging from inactive or overactive users, to accuracy of tweets themselves. Other areas of improvement that are not related to data mining is the optimization of the LSTM model or to forego LSTM altogether and use a different machine learning model. Additional refinement and modifications to the model is further needed.

References

[1] J. D. Hamilton, "Historical Oil Shocks," National Bureau of Economic Research, Working Paper 16790, Feb. 2011. [Online]. Available: https://www.nber.org/papers/w16790

[2] R. a. Markets, "Global $7425.02 Billion Oil and Gas Markets, 2015-2020, 2020-2025F, 2030F," Mar. 2021. [Online]. Available: https://www.globenewswire.com/news-release/2021/03/04/2187025/28124/en/Global-7425-02-Billion-Oil-and-Gas-Markets-2015-2020-2020-2025F-2030F.html

[3] ISPI, “Saudi Arabia’s Oil Dependence: Challenges Ahead,” Apr. 2016. [Online]. Available: https://www.ispionline.it/it/pubblicazione/saudi-arabias-oil-dependence-challenges-ahead-14997

[4] Organisation for Economic Co-operation and Development and Organisation for Economic Co-operation and Development, *Norway 2011*. Paris: OECD, 2011, oCLC: 1159660348.

[5] “EU imports of energy products - recent developments.” [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_imports_of_energy_products_-_recent_developments

[6] “What Is Oil Used For | Uses for Oil and Petroleum Products [Online]. Available: https://www.capp.ca/oil/uses-for-oil/

[7] A. Salzman, “Oil’s Spike Could Bring a Recession, History Shows.” [Online]. Available: https://www.barrons.com/articles/oil-prices-are-we-in-a-recession-51646240117

[8] “The Pioneers of Technical Analysis.” [Online]. Available: https://www.investopedia.com/articles/financial-theory/10/pioneers-technical-analysis.asp

[9] “What Is Technical Analysis?” [Online]. Available: https://www.investopedia.com/terms/t/technicalanalysis.asp

[10] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187775031100007X

[11] O. Sattarov, H. S. Jeon, R. Oh, and J. D. Lee, “Forecasting bitcoin price fluctuation by twitter sentiment analysis,” in *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, 2020, pp. 1–4.

[12] “Tutorial: Quickstart — TextBlob 0.16.0 documentation.” [Online]. Available: https://textblob.readthedocs.io/en/dev/quickstart.html

[13] “NLP For Beginners | Text Classification Using TextBlob,” Feb. 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

[14] M. Oussalah and A. Zaidi, “Forecasting weekly crude oil using twitter sentiment of u.s. foreign policy and oil companies data,” in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 201–208.

[15] Z. A. Sadik, P. M. Date, and G. Mitra, “Forecasting crude oil futures prices using global macroeconomic news sentiment,” *IMA Journal of Management Mathematics*, vol. 31, no. 2, pp. 191–215, 2020.

[16] M. Elshendy, A. F. Colladon, E. Battistoni, and P. A. Gloor, “Using four different online media sources to forecast the crude oil price,” *Journal of Information Science*, vol. 44, no. 3, pp. 408–421, Jun. 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0165551517698298

[17] “Papers with Code - Sentiment Analysis of Twitter Data for Predicting Stock Market Movements.” [Online]. Available: https://paperswithcode.com/paper/sentiment-analysis-of-twitter-data-for

[18] “Sentiment analysis of twitter data for predicting movement in stock price of apple inc. (aapl).” [Online]. Available: https://github.com/you915/Sentiment-Analysis-of-Twitter-Data-for-predicting-Apple-stock-price

[19] X. Li, W. Shang, and S. Wang, “Text-based crude oil price forecasting: A deep learning approach,” *International Journal of Forecasting*, vol. 35, no. 4, pp. 1548–1560, 2019.

[20] B. Wu, L. Wang, S. Wang, and Y.-R. Zeng, “Forecasting the us oil markets based on social media information during the covid-19 pandemic,” *Energy*, vol. 226, p. 120403, 2021.

[21] Z. Hu, “Crude oil price prediction using ceemdan and lstm-attention with news sentiment index,” *Oil & Gas Science and Technology–Revue d’IFP Energies nouvelles*, vol. 76, p. 28, 2021.

[22] “The Number of tweets per day in 2020,” Dec. 2019. [Online]. Available: https://www.dsayce.com/social-media/tweets-day/

[23] “Twitter: monthly active users worldwide.” [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[24] “GET /2/tweets/search/all.” [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

[25] “Rate limits.” [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/rate-limits

[26] “datetime — Basic date and time types — Python 3.10.4 documentation.” [Online]. Available: https://docs.python.org/3/library/datetime.html

[27] “re — Regular expression operations — Python 3.10.4 documentation.” [Online]. Available: https://docs.python.org/3/library/re.html

[28] “Crude Oil WTI Futures Historical Prices.” [Online]. Available: https://www.investing.com/commodities/crude-oil-historical-data

[29] “The Sequential model | TensorFlow Core.” [Online]. Available: https://www.tensorflow.org/guide/keras/sequential$_model$

[30] “sklearn.preprocessing.MinMaxScaler.” [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html