Title
Forecasting Oil Price Using Time Series Methods and Sentiment Analysis

Author
Lee, Shu Zhen

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Forecasting Oil Price Using Time Series Methods and Sentiment Analysis

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Shu Zhen Lee

2022

ABSTRACT OF THE THESIS

Forecasting Oil Price Using Time Series Methods and Sentiment Analysis

by

Shu Zhen Lee
Master of Applied Statistics
University of California, Los Angeles, 2022
Professor Frederic Paik Schoenberg, Chair

This study aims to predict West Texas Intermediate (WTI) crude oil spot price using ARMA models and sentiment analysis. Market sentiment is quantified using data from Twitter. Overall, four analyses are presented: 1) Baseline Model, 2) Lagged Regression, 3) AR(1), and 4) AR(1) + Sentiment. The baseline model simply uses the prior month's average as the current month's forecast. The lagged regression model uses Ordinary Least Squares (OLS) to regress the one-month lagged price on current price. The AR(1) model analyzes the monthly percent change and the AR(1) + Sentiment model adds market sentiment an additional predictor to the AR(1) model. Results indicate that an AR(1) + Sentiment is the best model and decreases the RMSE (as small RMSE is preferred) by 12.5% compared to the baseline model. However, the RMSE is quite large ($6.498/bbl) because the model fails to predict changes in trends and large jumps in price. Future work should mainly focus on improving these two items.

The thesis of Shu Zhen Lee is approved.

Hongquan Xu

Mark Stephen Handcock

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2022

To Kyle and Cooper

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Crude oil is a complex and vital component of the global economy. Its price is influenced by a variety of factors, such as exports, imports, global consumption, global production, international affairs, and Organization of the Petroleum Exporting Countries (OPEC) decisions. OPEC is a cartel protected by international law and greatly influences the price of oil. Corporations and governments that are exposed to crude oil risk have large incentives to understand and predict crude oil price. Forecasts can help companies and governments make short-term trades, plan long-term business decisions, and hedge risk.

Economic theory suggests that efficient markets follow a random walk [1]. However, crude oil markets have been extensively analyzed in literature with conflicting findings. As highlighted in Mishra (2016), some researchers find that crude oil prices revert to the mean, and are therefore predictable. Others find that crude oil prices follow a random walk pattern (non-stationary) and cannot be predicted [5]. This paper evaluates the stationarity of crude oil price and presents univariate time series analysis using ARMA models. As most crude oil volume is traded one or two months in advance, this paper focuses on monthly average forecasts. Specifically, the focus is to predict West Texas Intermediate (WTI) crude oil spot price one month into the future. WTI crude is the US crude oil benchmark, with the physical trading hub located in Cushing, Oklahoma.

Mathematical models assume players have perfect information and make logical and ra-

tional decisions based off that information. In reality, players have nowhere near perfect information. Instead, markets are driven by people's interpretation of events. For example, on April 20, 2020, the early days of the COVID-19 pandemic, WTI prices turned negative (-$37/bbl) for the first time in history [4]. Traders did not know how deadly or disruptive coronavirus would be, but collectively feared the pandemic would lead to significant global disruptions. Of course, COVID-19 did cause widespread disruptions, but not to the extent that crude oil prices should remain negative. Prices quickly pushed back into positive territory and remained in the $30-$40/bbl range for the majority of 2020. People's perception greatly overstated the risk to then downside during the early days of the COVID-19 pandemic.

Market sentiment and perception of current events are important factors to consider when predicting oil price, as illustrated by Zaidi and Oussalah (2018) and Zhao et al. (2019). However, market sentiment is difficult to measure and is insufficiently captured by traditional economic data. Therefore, unconventional data sources, such as social media, are utilized. Data is collected from Twitter because it is rich with information on people's opinions on the oil market. Collectively, tweets can provide a picture of market sentiment over time, which can then be used as an additional predictor in Auto-Regressive Moving Average (ARMA) models.

Overall, four analyses are presented: 1) Baseline Model, 2) Lagged Regression, 3) AR(1), and 4) AR(1) + Sentiment. The baseline model simply uses the prior month's average as the current month's forecast. The lagged regression model uses Ordinary Least Squares (OLS) to regress the one-month lagged price on current price. The AR(1) model analyzes the monthly percent change and the AR(1) + Sentiment model adds market sentiment an additional predictor to the AR(1) model.

1.2 Literature Review

Mishra (2016) questions whether commodity spot prices can be predicted. He employs unit root tests to determine whether natural gas prices follow a random walk or are mean reverting and hence predictable. Ultimately, he find that if heteroskedasticity and structural breaks are accounted for, prices are mean reverting and can be predicted. In his own literature review he highlights the fact that there has been extensive research on this topic with conflicting findings:

> "The results from existing studies have been mixed. Some studies have concluded that spot and/or futures energy prices are mean reverting and, hence, predictable (Elder and Serletis, 2008; Lee et al., 2006; Lee and Lee, 2009; Sadorsky, 1999; Serletis, 1992). Other studies have concluded that spot and/or futures energy prices are non-stationary or persistent or find mixed evidence that spot and futures prices are stationarity (Barros et al., 2014; Ghoshray and Johnson, 2010; Maslyuk and Smyth, 2008; Ozdemir et al., 2013; Pindyck, 1999; Presno et al., 2014)."

A recent study by Zaidi and Oussalah (2018) evaluates the sentiment of Twitter data on crude oil prices. The authors specifically focus on US foreign policy tweets and oil company tweets. The paper presents forecasts of weekly WTI crude price using Support Vector Machine (SVM), Naïve Bayes and Multi-layer perception (ML). Instead of predicting a continuous price variable, they choose to predict a categorical variable, either an "Increase" or "Decrease" in price [14].

This paper is different from Zaidi and Oussalah's in multiple ways. First, in addition to the directional categorization, this paper analyzes and predicts a continuous price variable in order to quantify the forecast. Second, whereas Zaidi and Oussalah's methodology used specific Twitter accounts regarding US foreign policy and oil companies, this paper analyzes tweets based on hashtags and therefore the tweets can stem from a wide variety of accounts.

In a separate study, Zhao et al. (2019) uses text data scraped from reliable news sources like Reuters and United Press International, Inc. (UPI) to predict Brent oil price. Brent, the European counterpart to WTI, is a similar crude oil benchmark. To find specific articles to scrape, the authors search for phrases such as "oil price", "OPEC", "Chevron", "Mobil", and "oil market". The authors use a method called VADER to conduct sentiment analysis. VADER splits the sentiment of each article into 4 measures: compound score, negative score, neutral score, and positive score. From there, Ridge, Lasso, Support Vector Regression (SVR), Back Propagation Neural Network (BPNN), and Random Forest (RF) are used as forecasting methods. The best results come from Lasso and Ridge although there are concerns with linearity. Findings demonstrate that adding the sentiments as predictors improves the results of the models. The root mean squared error (RMSE) decreases by about 0.2 and the error variance (used to measure stability of the results) also decreases by 0.2. This indicates a significant improvement in accuracy and stability [15].

CHAPTER 2

Data

2.1 Data Collection

2.1.1 Price Data

West Texas Intermediate (WTI) price data comes from the US Energy Information Administration (EIA) [7]. Specifically, it is the New York Mercantile Exchange (NYMEX) front month (CLC1) contract daily close price. WTI data goes back to 1983, but the analysis focuses on 2014 to 2022. The quality of WTI is as follows: "Specific domestic crudes with 0.42% sulfur by weight or less, not less than 37° API gravity nor more than 42° API gravity. The following domestic crude streams are deliverable: West Texas Intermediate, Low Sweet Mix, New Mexican Sweet, North Texas Sweet, Oklahoma Sweet, South Texas Sweet [11].".

It is important to note that crude oil can vary in quality and therefore in price. Different crude grades can have vastly different markets and pricing structures. For example, Alaskan North Slope (ANS) crude is typically traded along the US West Coast while Saudi Arabian crude oil is traded across the globe. Therefore, the methodology and results presented in this paper may not hold for other types of crude oil.

2.1.2 Twitter Data

Twitter data is collected through the Twitter API with special academic access [12]. Special academic access allows users to pull all tweets beginning in 2006. Over a million English

tweets from January 2014 to February 2022 are pulled with the hashtags #crudeoil, #WTI, #OPEC, #oilprices, #oott (organization of oil trading tweets). Retweets are excluded. Crude oil tweets with these hashtags may contain information about opinions and emotions of market participants.

The Twitter API is accessed through a Python script that continuously pulled 100 tweets at a time (up to 500 per day) over a specified date range. Each call to the API maxes out at 100 tweets, hence why only 100 tweets at a time can be pulled. The cap of 500 tweets a day was chosen because it is large enough to capture what occurred in the market that day, but also small enough to pull the data relatively quickly. Some example tweets are provided below.

- "#WTI Closing long position from low of the day. Happy with small gain.. will renter again."
- "#WTI #Brent EIA weekly inventory numbers very bullish."
- "#WTI crude higher on brighter demand prospects but now faces key resistance levels ahead http://t.co/2KW8hPFtF3 ÎR"
- "Short size #WTI crude oil avg 5485. #OOTT #Brent"
- "Central #banks loosening their grip on markets #OOTT #energy https://t.co/FghxE1nFF5 via @WSJ"

2.2 Twitter Data Cleaning

Tweets are textually messy, containing hyperlinks, punctuation, hashtags, etc. that make it difficult to perform text and sentiment analysis. Therefore, tweets must be cleaned of all unnecessary text prior to analysis. First, all hyperlinks, carriage returns, line breaks, and "&" are removed. Then, the tweets are tokenized using the R library tidytext.

Tokenization is a process that splits each tweet into single words, automatically makes the all letters lowercase, and removes punctuation. For example, if the tweet is "#WTI #Brent EIA weekly inventory numbers very bullish.", tokenization will turn this tweet into eight different records: 1) wti, 2) brent, 3) eia, 4) weekly, 5) inventory), 6) numbers, 7) very, 8) bullish. Then, regular expression ("[a-z']+") is used to remove numbers. Finally, stop words are removed, again using the R library tidytext. Stop words are common words in the English language such as "the", "and", and "this". These words serve as grammatical structure but do not invoke sentiment, emotions, or feelings, and are therefore unimportant to sentiment analysis.

Figure 2.1 presents a word cloud for tweets collected for 2017. It gives a basic idea of words that appear in the tweets. Specifically, 2017 is chosen as an example because of particular OPEC decisions that greatly influenced market sentiment. As a result, the word cloud shows significant discussion surrounding "saudi", "output", "cuts", "inventories" which is consistent with the important topics and main drivers of price changes in 2017, further explained in the next section.

Figure 2.1: Word Cloud of 2017 Crude Oil Tweets

2.3 Data Construction

2.3.1 Calculating Daily Sentiment of the Tweets

Sentiments are measured using Afinn scores, which were developed by Finn Arup Nielsen from 2009 to 2016. Each word is assigned an integer value from -5 to +5 depending on how negative or positive a word is. For example, the word "downside" is assigned a value of -2, the word "cut" is assigned a value of -1, and the word "gains" is assigned a value of +2 [6]. Some words do not have an associated value in the Afinn dictionary and therefore do not contribute to the overall sentiment score.

The tokenized version of the tweets include one record per word in a tweet. The Afinn scores are merged onto the tokenized dataset by word. The sentiments are then averaged for each individual tweet, giving each tweet a sentiment score. Finally, the scores are averaged over each month, creating 99 sentiment values, one for each month between January 2014 and February 2022. The monthly average sentiment is plotted below in Figure 2.2.

2.3.2 Interpreting Market Sentiment

From 2014 through 2022, market sentiment is the most negative from December 2015 to January 2016. Due to improvements in North American shale production, crude oil inventories had been steadily rising from mid 2014 through 2015 thus creating an oversupply in the market. As oversupply fears accelerated, traders' views turned bearish, resulting in negative market sentiment. To correct the market, OPEC agreed on November 30, 2016 to cut production by 1.2 MBD (millions of barrels per day) for six months beginning in January 2017 [9]. Surprisingly, market sentiment for the first half to 2017 remains negative, suggesting that the market believed the OPEC production cuts were not enough to fix the oversupply issue. OPEC met again on May 25, 2017 and agreed to extend production cuts by an additional nine months beginning July 2017 [8]. Following this decision, market sentiment

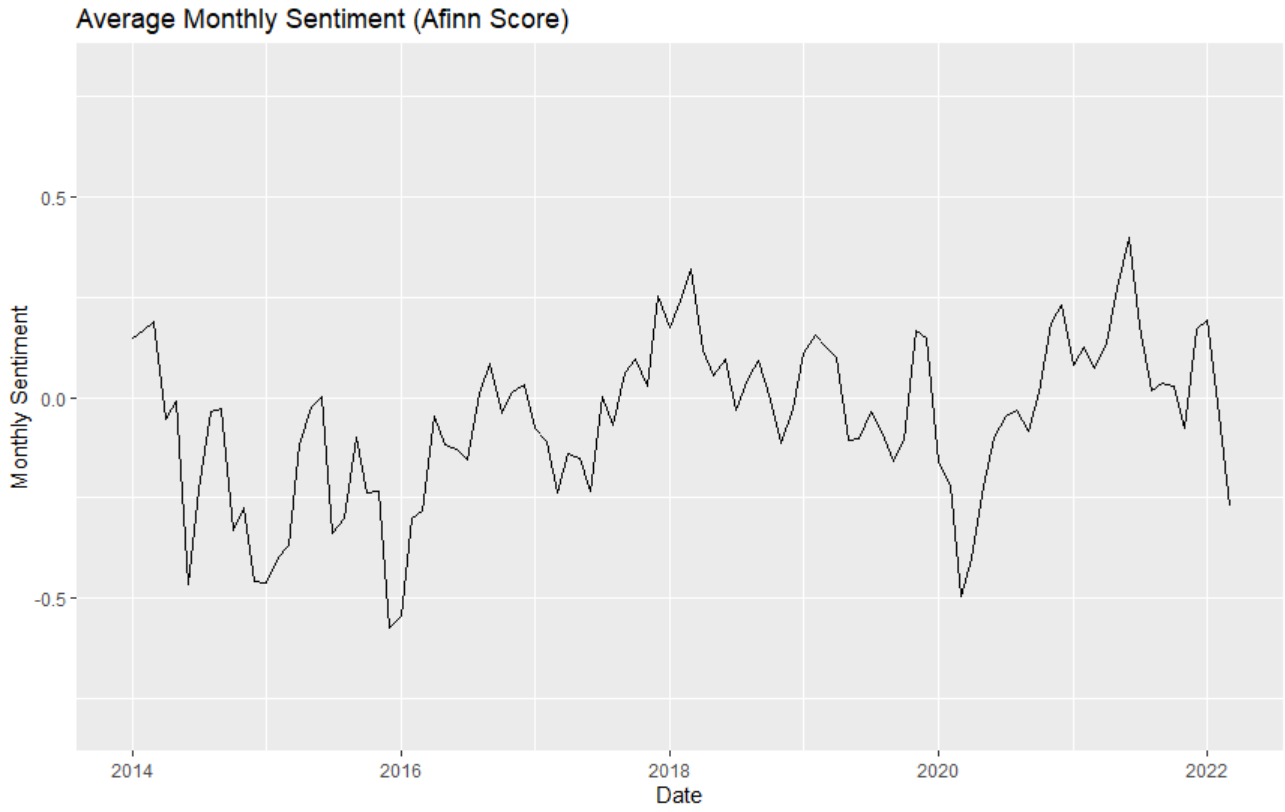

Figure 2.2: Monthly Average Sentiment - Calculated Using Afinn Scores

shifts into the positive territory and continues to remain positive until the fourth quarter of 2018. Market sentiment remains fairly neutral through 2019 and then turns abruptly negative as COVID-19 spreads globally, essentially eliminating demand for petroleum products. Sentiment rebounds as the world slowly reopens but then crashes again when Russia invades Ukraine in early 2022. Russia is a major oil player and the loss of Russian oil (due to sanctions) cannot be immediately replaced in the market.

The strongly negative sentiments in 2016 and 2020 are based on fears that oil price will decrease. However, the negative sentiment surrounding Russia's invasion of Ukraine in 2022 is based on fears of prices increasing. This shows that there are instances when the same sentiment value describes opposing price movements. Predicting 2022 prices using a model trained on sentiments from 2016 to 2020 data may not perform well.

2.3.3 Creating Monthly Averages and Percent Change

Daily WTI price data is averaged over each month to create the monthly average price. Then, the monthly average price is lagged by one month to create a predictor variable for the baseline and lagged regression models. Most importantly, the monthly percent change is calculated to create a stationary time series for the AR(1) models as shown in Figure 2.3 below. The monthly percent change typically lies in the range of -12.5% to +12.5%. However, from February 2020 to June 2020 WTI price exhibits strong volatility, stemming from demand shocks cause by COVID-19. In this time period, the percent change in WTI price is as low as -39.8% and as high as +48.9%.

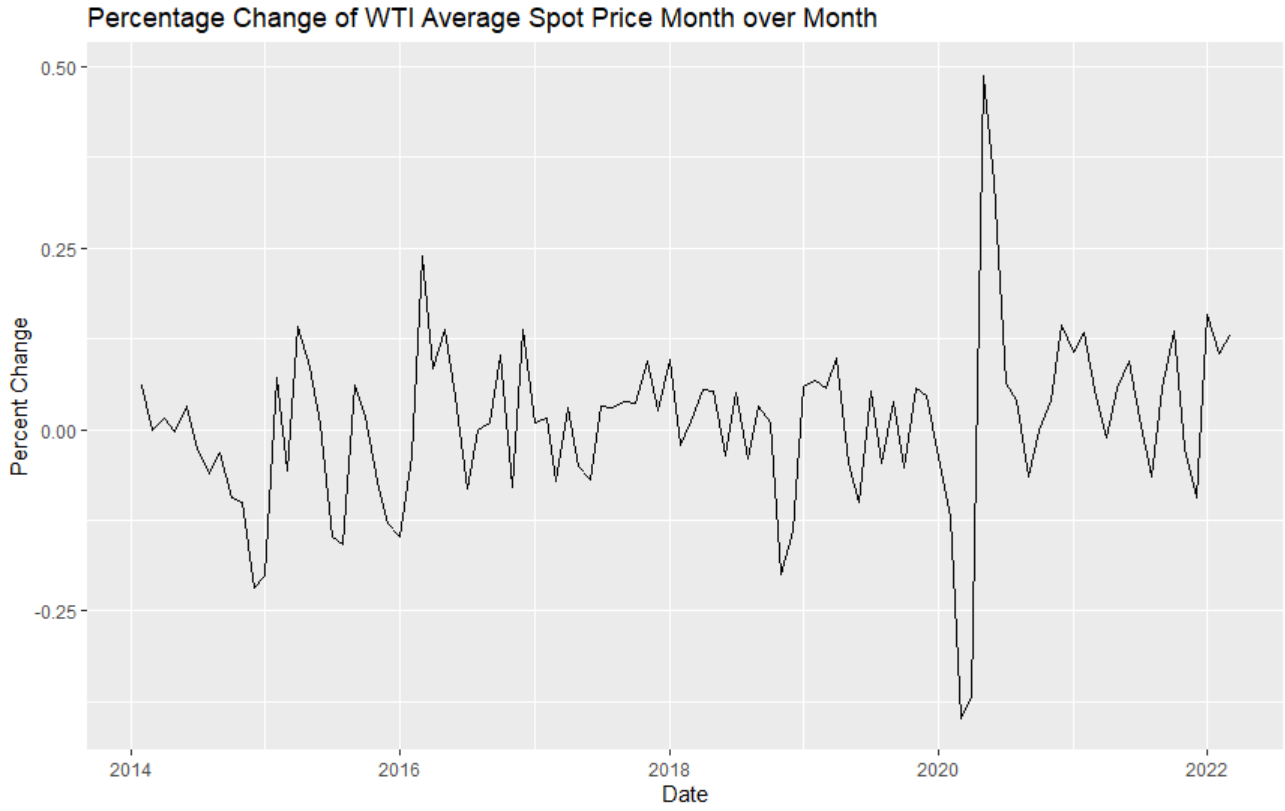

Figure 2.3: Monthly Percent Change of WTI Spot Oil Price

Similarly, daily Twitter sentiment data is converted to a monthly average and then merged to the price data by month. The sentiment is lagged by one month (i.e. January

sentiment will be attached to February percent change in price), for use as an additional predictor in the AR(1) model. Finally, the data is split into training and testing datasets. The training data consists of all data before January 2020 and the testing dataset includes January 2020 and later.

CHAPTER 3

Methodology

3.1 Baseline Method and Lagged Regression

A baseline model is created under the assumption that the average monthly price of month t will be the average monthly price of month $t + 1$. Simply put, this model assumes no monthly change in WTI price. All other models are compared to this model as it is the simplest model to create. If a model does not improve upon the baseline model, it is not considered viable. Next, another simple model is created using ordinary least squares (OLS) regression. The dependent variable is the monthly average price and the only predictor is the one-month lag of the monthly average price. From here on, this model is called the "lagged regression model".

3.2 Stationary Time Series

The next part of the analysis is based on ARMA models. ARMA models require a stationary time series. A stationary time series has no obvious trend and has constant variance [10]. The Augmented Dickey-Fuller (ADF) test is a unit root test that evaluates whether a time series is stationary or not. If the time series contains a unit root, the time series is not stationary. In an ADF test, the alternative hypothesis is that the time series does not contain a unit root, and is therefore stationary [2].

The ADF test for the outright monthly WTI price has a p-value of 0.08, therefore the

null hypothesis is not rejected, suggesting the outright WTI price is not stationary. An ADF test on the monthly percent change in WTI price has a p-value of 0.01. Consequently, the null hypothesis is rejected, demonstrating that the percent change in price is approximately stationary.

3.3 Autocorrelation

It is then necessary to evaluate which past periods, s, influence the price of the current period, t. The autocorrelation function (ACF) measures "the linear predictability of the series at time t, say x_t, using only the value x_s" [10]. Equation 3.1 below defines the autocovariance function γ, as the covariance between past time period, s, with the current time period, t. Equation 3.2 shows the formula for the autocorrelation of s and t, where $-1 \leq \rho(s,t) \leq +1$.

$$\gamma(s,t) = cov(x_s, x_t) \tag{3.1}$$

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \tag{3.2}$$

Because monthly percent change is approximately stationary, the autocovariance $\gamma(s,t)$ will only depend on the difference between s and t, $|s-t|$. Equations 3.1 and 3.2 can be re-written as Equations 3.3 and 3.4 where h represents the time lag. For this analysis, $h = 1$ is a one month lag.

$$\gamma(h) = cov(x_{t+h}, x_t) \tag{3.3}$$

$$\rho(t,h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} \tag{3.4}$$

Below, Figure 3.1 plots the monthly average training data over time. It also presents the autocorrelation function (ACF) and partial autocorrelation function (PACF) charts of the monthly average training data. The blue dashed lines in the ACF and PACF plots indicate the 95% confidence interval bands. ACFs and PACFs outside of the blue dashed

lines indicate that that particular lag is significant in predicting the percent change at time t. The only significant lag in the ACF plot is the first lag, $h = 1$. Similarly, the only significant lag in the PACF plot is the first lag, $h = 1$. Given these results, either ARMA(1,1), AR(1), or MA(1) is likely to be the best model.

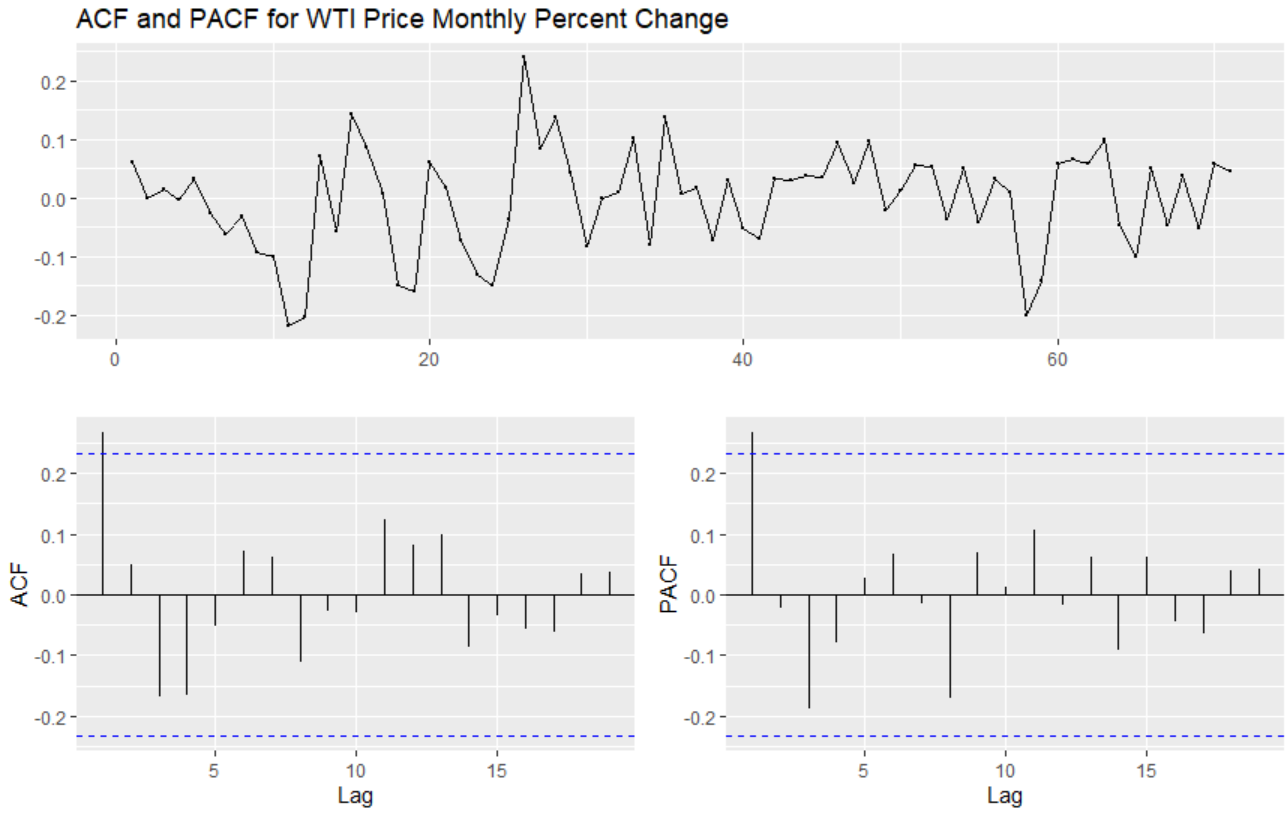

Figure 3.1: ACF and PACF for WTI Price Monthly Percent Change

3.4 ARMA Modeling and Residual Analysis

The auto.arima function in R will identify the best ARMA model to use based on Akaike information criterion (AIC). Results for AR(1), MA(1), and ARMA(1,1) are presented below in Table 3.1. The results suggest that AR(1) is the best model to use as the AIC score (-146.72) is the lowest of all tested models. However, the MA(1) and ARMA(1,1) models have

AIC scores that are similar. For future work, evaluating these additional models may result in useful predictions. However, the rest of the report focuses on the AR(1) model.

Table 3.1: AIC for Models on the Training Data

Model	**AIC**
AR(1)	-146.72
MA(1)	-144.24
ARMA(1,1)	-142.75

After fitting an AR(1) model to the training data, the next step is to confirm whether there are any additional significant autocorrelations. The Ljung-Box test evaluates whether any groups of autocorrelations of a time series are different from zero [10]. The null hypothesis is that all autocorrelations are not significantly different from zero, thus the model is adequate. The alternative hypothesis states that at least one autocorrelation is significantly different from zero. For the AR(1) model, the Ljung-Box Q statistic is $Q = 7.67$ for 12 lags with p-value = 0.63. Hence, the null hypothesis is rejected and the AR(1) model is deemed adequate.

The ACF plot in Figure 3.2 shows similar results to the Ljung-Box test. The plot shows that none of the autocorrelations cross the blue 95% confidence interval lines, suggesting there are no further significant autocorrelations. The bottom-right plot shows a histogram of the residuals which has two distinct peaks, revealing possible issues with normality. The top plot displays the residuals over time and highlights no obvious pattern, however the residuals are further analyzed for cyclical patterns in the next section.

3.5 Spectral Analysis

Thus far, price is analyzed in the time domain. However, with spectral analysis, time series can also be analyzed in the frequency domain which expresses the series in terms of underlying periodic components. For example, it is known that US gasoline prices exhibit

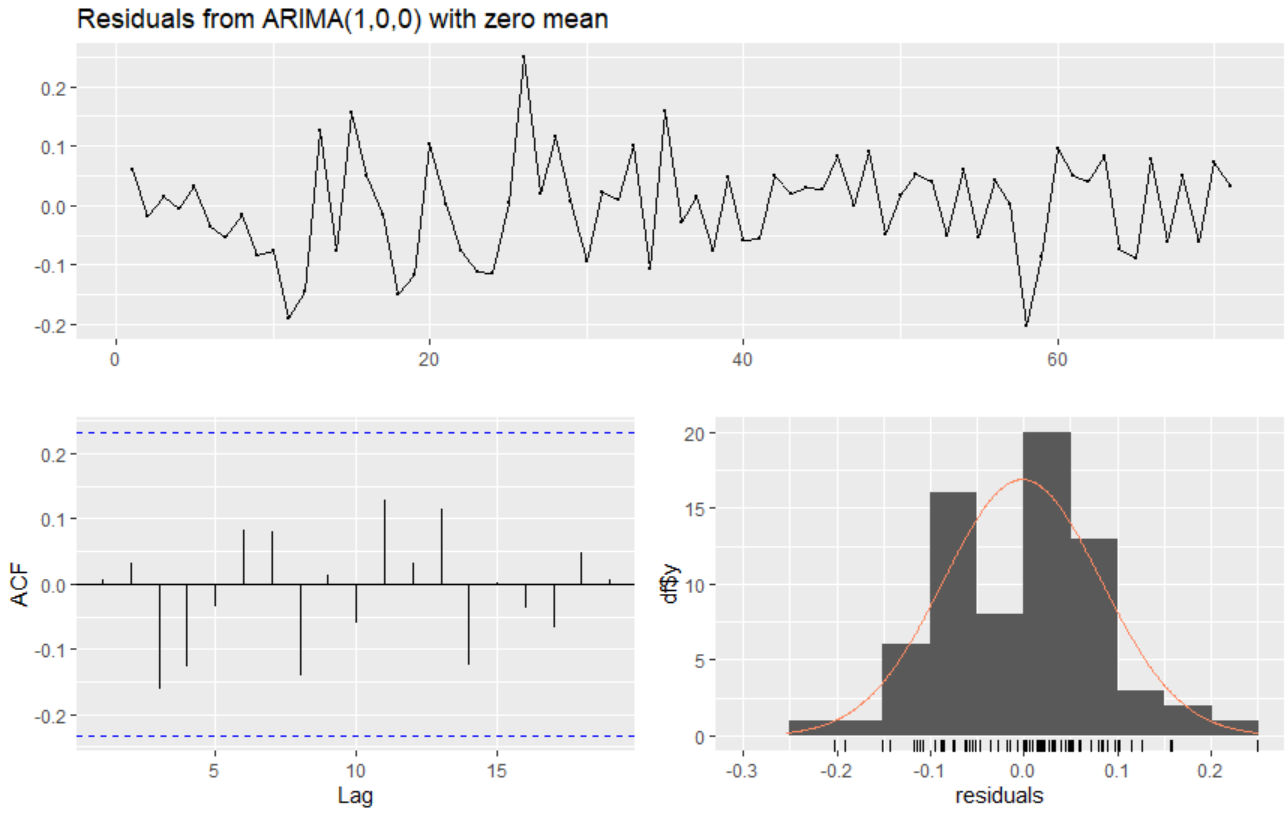

Figure 3.2: AR(1) Residual Plots

annual seasonality, or one cycle every 12 months. Gasoline prices generally increase in the summer when people are on summer vacation, and fall again during the fall and winter months [13].

To evaluate whether WTI price also exhibits seasonality, the spectral density of the AR(1) residuals is analyzed. The autocovariance function shown in Equation 3.3 can be represented by Equation 3.5 where ω represents the frequency. This can then be converted into the spectral density (the discrete time Fourier transform of the autocovariance) using Equation 3.6.

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i\omega h} f(\omega) d\omega, h = 0, \pm 1, ... \tag{3.5}$$

$$f(\omega) = \sum_{-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}, -1/2 \leq \omega \leq 1/2 \tag{3.6}$$

Figure 3.3 below shows the spectral density graph of the AR(1) residuals. The x-axis is the frequency of the time series, in cycles per 12 month (annual) periods. The y-axis is the spectrum of the time series which indicates the power of different frequencies. If the residuals exhibit annual behavior, Figure 3.3 should show a peak shown at a frequency of 1. Instead, the spectral density graph presents a flat horizontal line, demonstrating that the residuals are likely white noise.

Additionally, Figure 3.4 plots the smoothed periodogram for the AR(1) residuals. Again, the x-axis is the frequency and the y-axis is the spectrum. Note the range of the y-axis is much narrower than that of Figure 3.3. If plotted on the same y-axis, the smoothed periodogram would also show a horizontal line. However, the periodogram may suggest weak cyclic behavior at frequencies of 2 and 5, corresponding to cycles every 6 months and 2.4 months respectively.

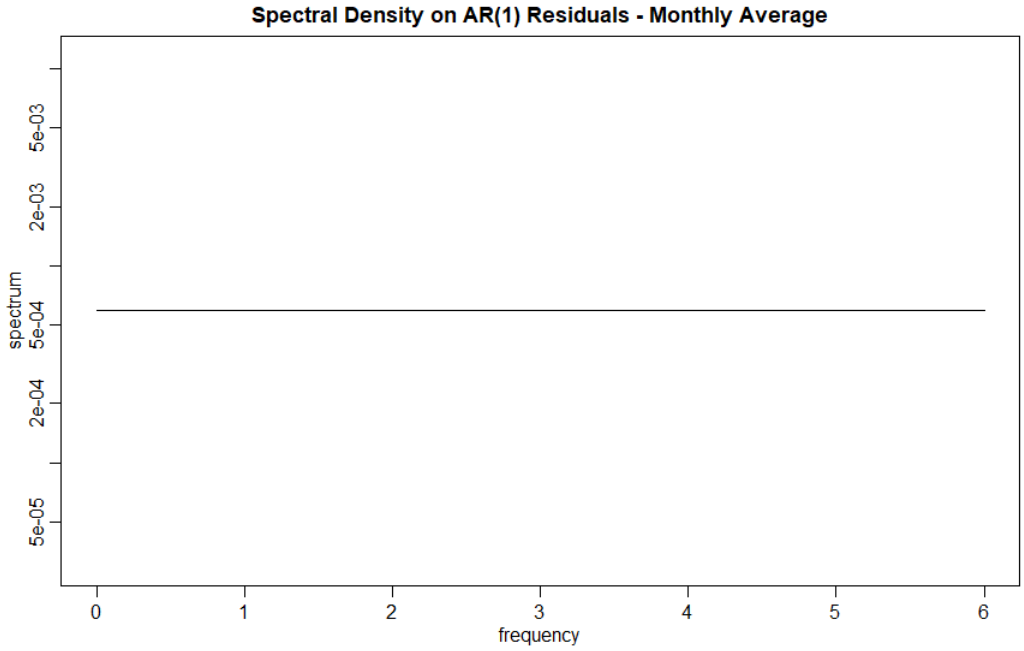

Figure 3.3: Spectral Graph of AR(1) Residuals

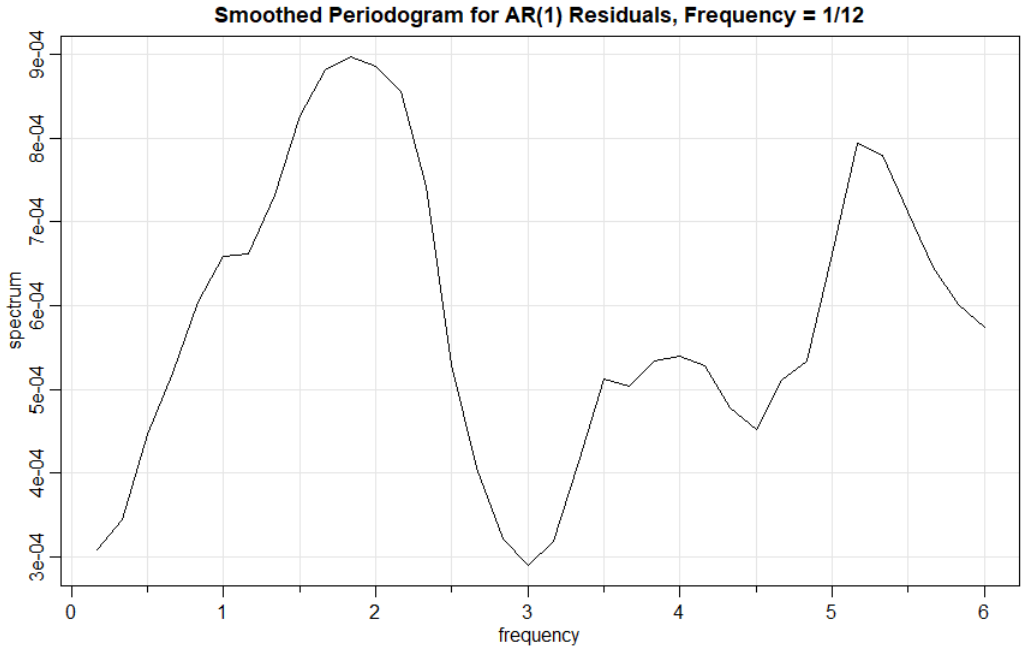

Figure 3.4: Smoothed Periodogram of the AR(1) Residuals

3.6 Adding Sentiments from Tweets

As discussed in detail in Chapter 2, the daily market sentiment is averaged over each month to create a monthly average. The monthly average market sentiment is merged onto the price data by month and then lagged by one month so the previous month's market sentiment is attached to the current month's price. Another AR(1) model is fit to the training data, including market sentiment as a predictor.

The residual analysis presents similar results to the original AR(1) model. The Ljung-Box test (as described in Section 3.4) has a p-value of 0.55, demonstrating that the other autocorrelations are not significantly different from zero, thus the AR(1) + Sentiment model is sufficient. Figure 3.5 shows the residuals over time (top), the ACF plot (bottom-left), and the histogram of the residuals (bottom-right). The ACF plot shows similar results to the Ljung-Box test as none of the autocorrelation bars cross the dotted blue 95% confidence interval lines. The AR(1) + Sentiment histogram only has one peak whereas the AR(1)

histogram (shown in Figure 3.2) had two distinct peaks, suggesting the AR(1) + Sentiment model may meet the normality condition better than the AR(1) model.

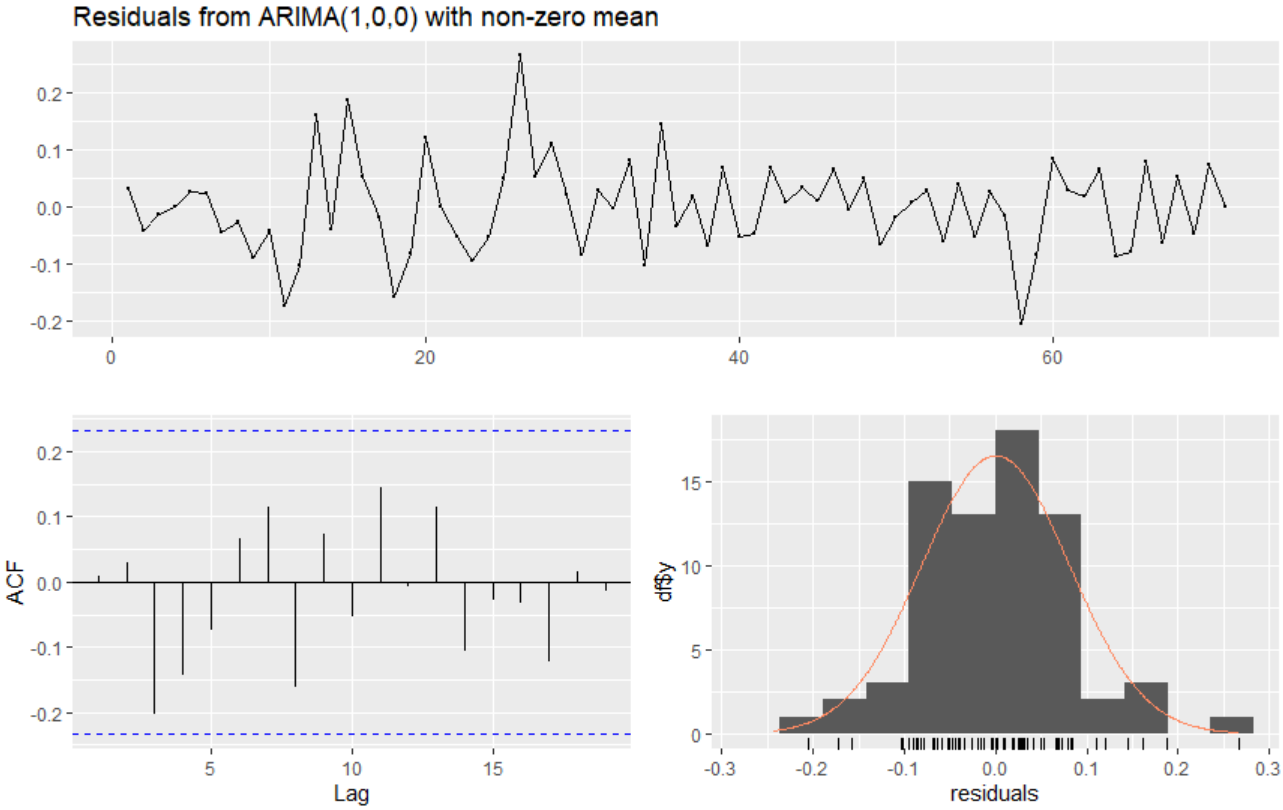

Figure 3.5: AR(1) + Sentiment Residual Plots

Figures 3.6 and 3.7 below show the spectral density plot and the smoothed periodogram plot respectively for the AR(1) + Sentiment residuals. The spectral plot again shows a horizontal line, highlighting no obvious seasonal pattern in the residuals. The periodogram would also show a horizontal line if plotted on the same y-axis as the spectral graph. However, when plotted with a narrower y-axis range, periodogram shows peaks at frequencies of 2 and 5, suggesting weak cyclic behavior every 6 and 2.4 months respectively. Thus, results imply that the AR(1) + Sentiment residuals are likely white noise.

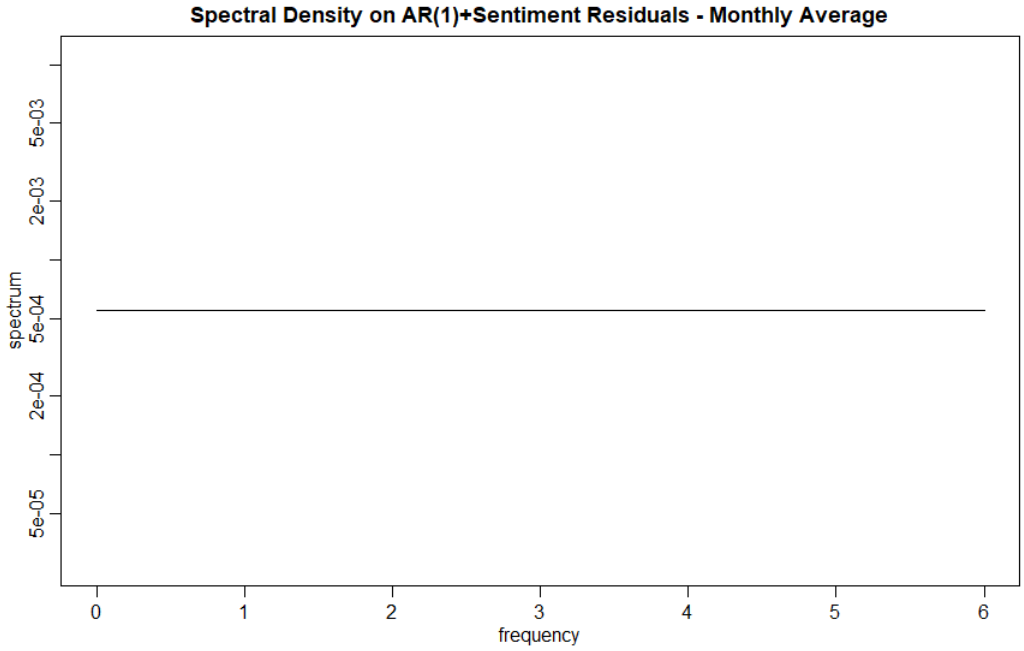

Figure 3.6: Spectral Graph of AR(1) + Sentiment Residuals

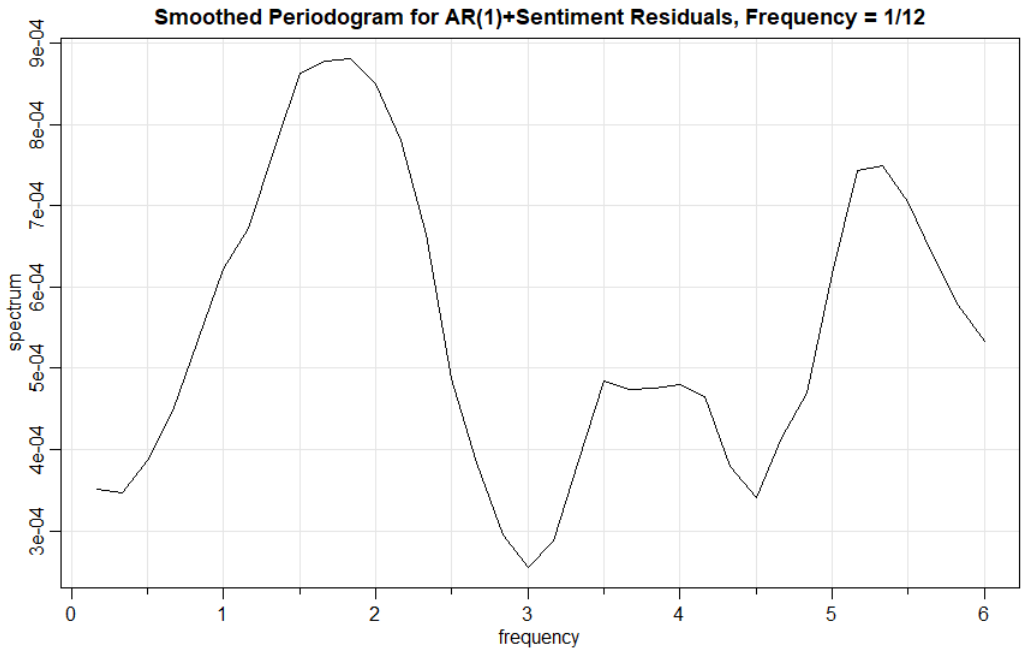

Figure 3.7: Smoothed Periodogram of the AR(1) + Sentiment Residuals

CHAPTER 4

Results

Thus far, four important things are established:

1. The ADF test indicates that monthly percent change is approximately stationary.
2. An AR(1) is the best ARMA model to model monthly percent change as it has the smallest AIC value.
3. An AR(1) adequately models the monthly percent change and has no further autocorrelations that are significant.
4. The spectral graphs indicates that AR(1) and AR(1) + Sentiment residuals are likely white noise, suggesting weak seasonality, if any.

From here, forecasts are created from January 2020 to February 2022 for the baseline, lagged regression, AR(1), and AR(1)+Sentiment models. These results are described in Table 4.1 below.

4.1 Forecasting

In a trading environment, the majority of oil is traded one month in advance. If traders can accurately forecast the price of the next month, they can adjust their trading strategy to buy or sell depending on their current position. Therefore, forecasts are conducted one month at a time and only predict one month out. For example, the forecast for January

2020 includes all data through December 2019 and the forecast for February 2020 includes all data through January 2020. Consequently, for each month a new model is created using the data from all past months. The same method is applied for the lagged regression model, AR(1) model, and AR(1) + Sentiment model.

4.2 Monthly Results

The results of the baseline model, lagged regression model, AR(1) model, and AR(1) + sentiment model are presented in Table 4.1. All months between January 2020 through March 2022 are forecasted, resulting in 27 months total. The second column in the table is the root mean squared error (RMSE). The third and fourth columns represent how often the AR(1) models predicts the direction (increase or decrease) of the price change correctly or incorrectly.

Table 4.1: Monthly Forecasting Results

Model	**RMSE $/bbl**	**Direction Correct %**	**Direction Incorrect %**
Baseline	7.430	-	-
Lagged Regression	7.702	-	-
AR(1)	6.790	63.0	37.0
AR(1) + Sentiment	6.498	74.1	25.9

The RMSE calcuates the average distance between the actual (y_i) and predicted ($\hat{y}_i$) values, as shown in Equation 4.1 below. For example, the baseline model has an RMSE of \$7.43/bbl which suggests the difference (error) between the average and predicted price is on average \$7.43/bbl. A smaller RMSE is ideal because indicates that the distance between the actual and predicted values is smaller and therefore the predictions are more accurate.

$$RMSE = \sqrt{\frac{\sum_i^n (y_i - \hat{y}_i)^2}{n}} \tag{4.1}$$

The lagged regression's RMSE is \$7.702/bbl, the largest out of the four models presented.

It performs 3.7% worse than the baseline model and is, therefore, not a model that would be practical to use in a real-world setting.

The AR(1) model predicts the direction of the price change correctly 63% of the time. The RMSE is $6.79/bbl, which improves upon the baseline model by 8.6%.

The AR(1) + Sentiment model is the best model out of all four presented in Table 4.1 because it has the smallest RMSE. The difference in RMSE between this and the baseline model is $0.932/bbl or a 12.5% improvement. Improving the model's RMSE by almost $1.00/bbl increases the confidence in the prediction. With stronger predictions, traders can increase profits (buying more now to sell later) or limit losses (selling now if they know prices will decline). This model also performs better with predicting the direction of the price movement. It predicts the direction correctly 74.1% of the time.

There is an important distinction betwen the lagged regression model and the AR(1) models. An AR(1) predicts percent changes and converts the percentages back into a dollar amount. On the other hand, the lagged regression predicts an outright price in $/bbl. As the results from the lagged regression are worse than the AR(1) models, this suggests that dependent variables that are stationary (like percent changes) are preferred. Predicting percent change and converting back to dollars results in a better forecast compared to predicting the price outright.

4.2.1 Time Series Graphs with Forecast

Forecasts (shown in red) are presented from January 2020 to February 2022 in the four figures below. Figure 4.1 shows the baseline model. Figure 4.2 shows the forecasted monthly prices using the lagged regression model. Figure 4.3 shows the forecasted monthly prices using the AR(1) model. Figure 4.4 shows the forecasted monthly prices using the AR(1) + Sentiment model.

Although the forecasts of all models look similar, the gap between the predicted vs actual

prices (i.e. the gap between the red and the black lines in the charts) are different. The lagged regression model in Figure 4.2 shows almost no overlap between the red and the black line. The AR(1) model in Figure 4.3 shows a much tighter prediction, where the forecasted values are much closer to the actual values. The AR(1) + Sentiment model in Figure 4.4 shows an even tighter prediction than the AR(1) model, particularly during the upward trend exhibited from November 2020 to March 2021. Adding market sentiment captures some variation in prices that helps adjust the prediction closer to the actual value. Overall, the AR(1) + Sentiment model performs the best out of the four presented. Although market sentiment takes a bit of effort to measure, it is worthwhile to do so when predicting WTI crude oil price.

In all four charts, the forecasts begin in 2020 and have to predict the large price downturn caused by the demand shock from the COVID-19 pandemic, as explained in Section 2.3.2. Because all four methods use information from the prior month's price, the general downward trend in early 2020 is captured by all models. In fact, all models perform well in capturing trends. If the price continuously increases or decreases for multiple periods, the forecast tends to follow the trend. Conversely, all models struggle to capture changes in the price trend. For example, beginning in January 2020 there followed three consecutive months of price decreases. However, from April 2020 ($19.16/bbl) to May 2020 ($28.53/bbl) this trend reverses and prices begin increasing. Erroneously, all models continued to forecast a price decrease instead of a price increase. With regards to the entire time period, the AR(1) + Sentiment model captures trend changes better than the pure AR(1) model, predicting 74.1% of the directions successfully vs only 63.0% in the AR(1) model.

Furthermore, even if the trend remains the same, the models have difficulty predicting large jumps in price. For example, from February 2020 to March 2020 the monthly average price fell from $50.54/bbl to $30.45/bbl, a drop of 39.75%. Although the downward trend, which began in December 2019, continued, the forecasts missed the mark by 58-66% as shown in Table 4.2. Similarly, there was a large price increase from February 2022 to March 2022.

The price increased from $91.63/bbl to $103.41/bbl, a 12.9% increase. The forecasts were substantially off, with errors between 8.2% and 13.2% as shown in Table 4.3. March 2020 and March 2022 (only 2 out of 27 (7.4%) months forecasted) make up a disproportionate amount of the error. Table 4.4 shows that these two months make up 14.6-17.5% of the RMSE. Improving predictions on trend reversals and large jumps in magnitude are two keys areas of further research.

Table 4.2: Forecasts for March 2020

Model	**Actual Price $/bbl**	**Predicted Price $/bbl**	**% Difference**
Baseline	30.45	50.54	65.98
Lagged Regression	30.45	50.52	65.91
AR(1)	30.45	48.70	59.93
AR(1) + Sentiment	30.45	48.01	57.67

Table 4.3: Forecasts for March 2022

Model	**Actual Price $/bbl**	**Predicted Price $/bbl**	**% Difference**
Baseline	103.41	91.63	-11.39
Lagged Regression	103.41	89.81	-13.15
AR(1)	103.41	94.94	-8.19
AR(1) + Sentiment	103.41	93.89	-9.21

Table 4.4: Monthly Results Removing March 2020 and March 2022

Model	**Orig RMSE $/bbl**	**New RMSE $/bbl**	**% Difference**
Baseline	7.430	6.159	17.11
Lagged Regression	7.702	6.358	17.45
AR(1)	6.790	5.799	14.59
AR(1) + Sentiment	6.498	5.444	16.22

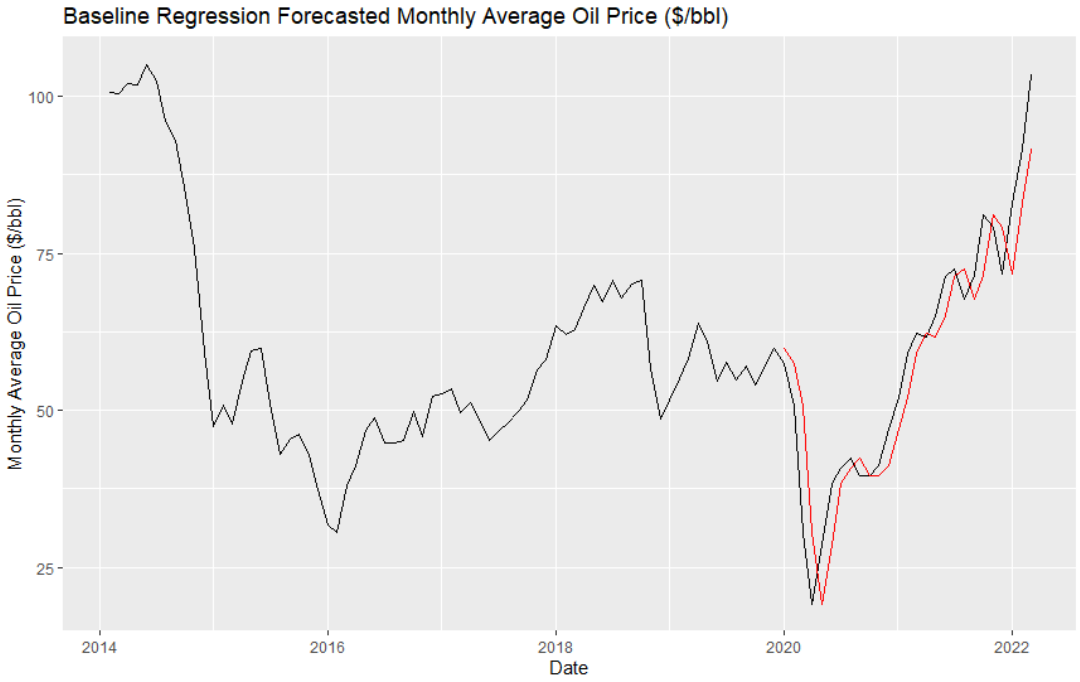

Figure 4.1: Monthly Forecast for Baseline Model

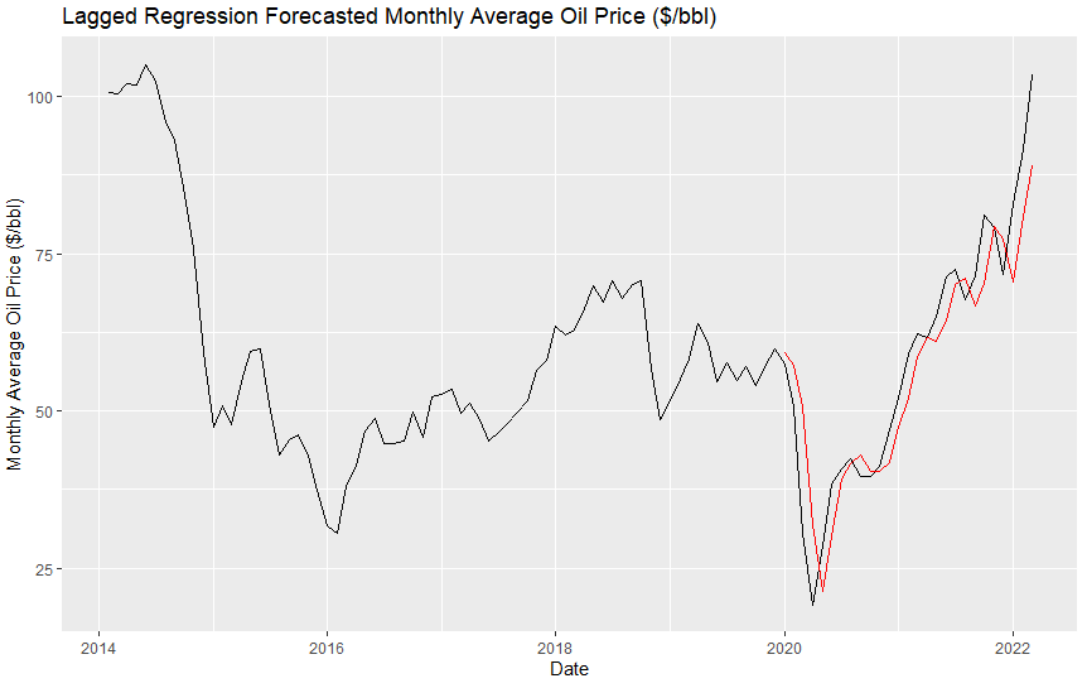

Figure 4.2: Monthly Forecast for Lagged Regression Model

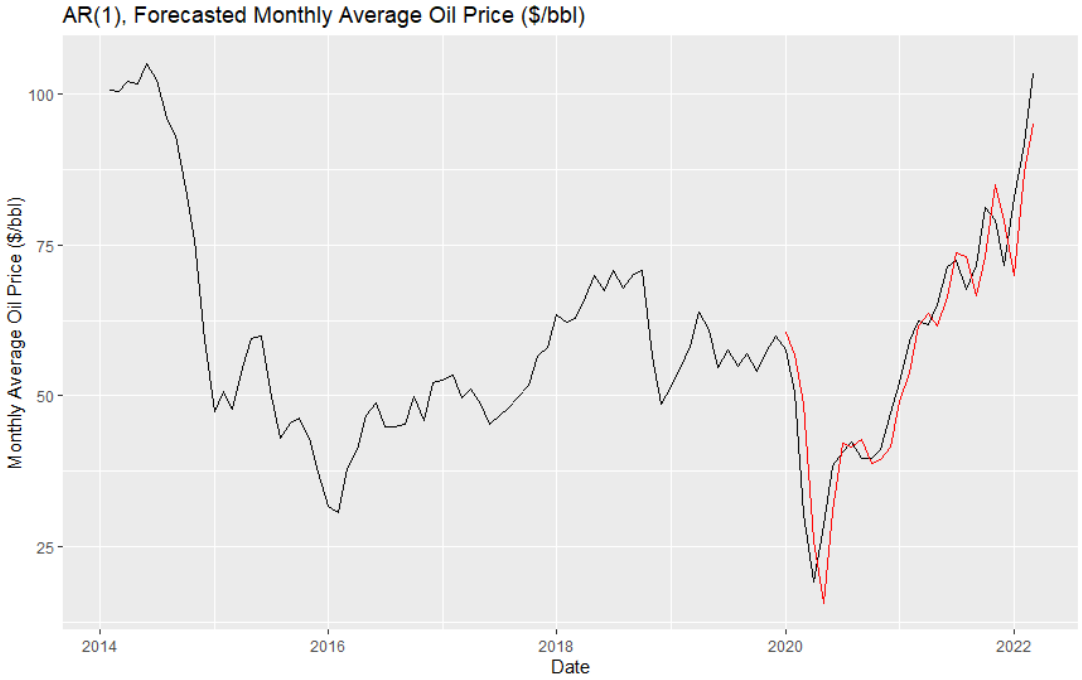

Figure 4.3: Monthly Forecast for AR(1) Model

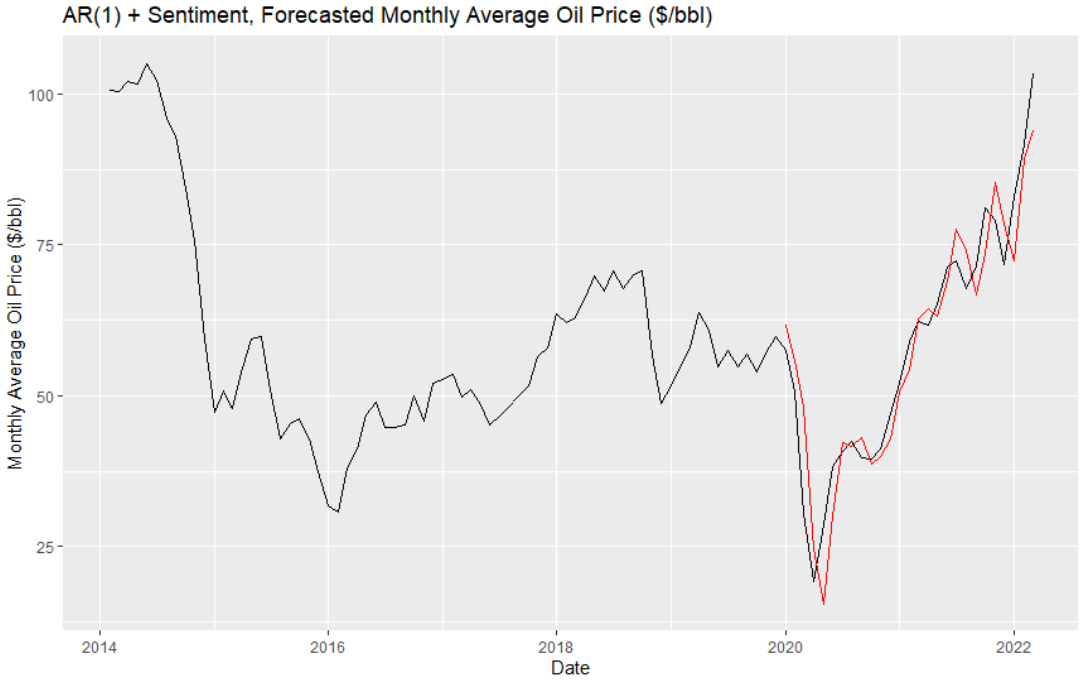

Figure 4.4: Monthly Forecast for AR(1) + Sentiment Model

CHAPTER 5

Conclusion

Chapter 1 introduced the overall problem and reviewed relevant literature. Chapter 2 discussed data and went into detail about how market sentiment is measured from Twitter data. Chapter 3 presented the methodology and evaluated different assumptions and criteria of ARMA models. An ARMA model must use a stationary time series. The ADF test, showed that the WTI monthly percent change is approximately stationary. The auto.arima function found that an AR(1) model fits the training data the best. Residual analysis confirmed that no further autocorrelations were considered significant. Spectral analysis showed that the residuals are likely white noise, suggesting weak seasonality in the WTI price, if any.

Chapter 4 presented the results from four different models, the baseline, lagged regression, AR(1), and AR(1) + Sentiment models. Of the four models presented, the AR(1) + Sentiment model best predicts the monthly average WTI price, improving performance by 12.5% compared to the baseline model. The market sentiment captured from the Twitter data is successful in improving the forecasts and is worthwhile to collect. The worst performing model is the lagged regression model, suggesting that going forward, percent changes in price should be modeled instead of the outright price.

5.1 Critique and Future Work

5.1.1 Potential Improvements

The methodology requires the price data for the entire month prior to the month of interest. Therefore, forecasts can only be made one month out. In practice, it may be useful to have stable forecasts multiple months into the future. In addition, the monthly time interval is the only time interval considered. Weekly, quarterly, or yearly intervals may be of interest to different parties.

Furthermore, as discussed in Chapter 4, the current method does a poor job of predicting changes in trends and large jumps in price. Accurate predictions of large price movements can potentially improve the model by 16.2% as shown in Table 4.4. Being able to predict changes in trends may improve the model even further.

For specific trading applications, instead of forecasting outright spot prices, forecasting specific NYMEX futures prices may help traders choose which specific contracts they want to enter or exit. Taking into account market structure (i.e. contango or backwardation) and actual trading costs like freight and taxes may also improve the model.

5.1.2 Investigate Other Techniques and Additional Predictors

The main method of this paper is an AR(1) model. To improve the robustness and confidence of the forecast, it would be interesting to evaluate whether different forecasting methods (i.e. classification, Support Vector Machine (SVM), Ridge Regression, etc.) converge to the same answer. Additional predictors could be added as well. Other features can be created from the Twitter data, such as weighting sentiments by the number of retweets. Moreover, a plethora of economic data (exports, imports, production, refinery runs, etc.) can be gathered from the Energy Information Administration's website that may be useful predictors as well.

5.1.3 Quantifying Sentiment and Topic Modeling

Sentiment can be measured multiple different ways and it would be interesting to evaluate the performance of different sentiment measures. In particular, Google's BERT (Bidirectional Encoder Representations from Transformers) algorithm uses the context of a body of text to evaluate sentiment, not just the words themselves, which may lead to a more accurate sentiment measure [3]. Another possible area of interest is topic modeling. Certain topics like "OPEC" or "shale" may appear more often when the market sentiment is bullish or bearish. It would be interesting to determine how adding topics as a feature improves the forecast, if at all.

5.1.4 Forecast Other Commodity Prices

To evaluate whether this methodology can be generalized, it should be extended to other commodities such as gasoline, diesel, natural gas, and other types of crude oil. The underlying tweets may have to be adjusted for different commodities. For example, the hashtags #gasoline or #diesel may be included for the respective commodities. Other hashtags like #oott or #OPEC may not influence other commodities as much as they influence crude oil. Furthermore, seasonality may play a bigger role in the gasoline and heating oil prices. More in depth spectral analysis is likely needed for those commodities. However, the structure of the overall methodology should remain relatively similar.

REFERENCES

[1] Baghestani, Hamid. "Inflation Expectations and Energy Price Forecasting." OPEC Energy Review, vol. 38, no. 1, Mar. 2014, pp 21–35. https://doi.org/10.1111/opec.12016.

[2] Baltagi, Badi H. Econometrics 5th edition, Berlin, Springer International Publishing, 2011.

[3] Devlin, Jacob et al. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Arxiv, 2018, pp. 1-16. https://doi.org/10.48550/arXiv.1810.04805.

[4] Irwin, Niel. "What The Negative Price Of Oil Is Telling Us (Published 2020)". New York Times, 2022, https://www.nytimes.com/2020/04/21/upshot/negative-oil-price.html. Accessed 12 May 2022.

[5] Mishra, Vinod, and Russell Smyth. "Are Natural Gas Spot and Futures Prices Predictable?" Economic Modelling, vol 54, Apr. 2016, pp. 178–186. https://doi.org/10.1016/j.econmod.2015.12.034.

[6] Nielsen, Finn Årup. "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs." Arxiv, 2011, pp. 1-6. https://doi.org/10.48550/arXiv.1103.2903.

[7] "NYMEX Futures Prices". US Energy Information Administration (EIA), EIA, https://www.eia.gov/dnav/pet/pet_pri_fut_s1_d.htm. Accessed 3 May. 2022.

[8] "OPEC : OPEC 172nd Meeting Concludes." Organization of the Petroleum Exporting Countries, www.opec.org/opec_web/en/press_room/4305.htm. Accessed 7 May 2022.

[9] "OPEC : OPEC and Non-OPEC Ministerial Meeting." Organization of the Petroleum Exporting Countries, www.opec.org/opec_web/en/press_room/3944.htm. Accessed 7 May 2022.

[10] Shumway, Robert H., and David S. Stoffer. Time Series Analysis and Its Applications with R Examples 3rd Edition, New York City, Springer International Publishing, 2017.

[11] "Table Definitions, Sources, and Explanatory Notes - Petroleum Prices". US Energy Information Administration (EIA), EIA, https://www.eia.gov/dnav/pet/TblDefs/pet_pri_fut_tbldef2.asp. Accessed 3 May. 2022.

[12] "Twitter API for Academic Research Products Twitter Developer Platform." Twitter, Twitter, https://developer.twitter.com/en/products/twitter-api/academic-research. Accessed 3 May. 2022.

[13] "U.S. Energy Information Administration - EIA - Independent Statistics and Analysis." Gasoline Price Fluctuations, U.S. Energy Information Administration (EIA), https://www.eia.gov/energyexplained/gasoline/price-fluctuations.php. Accessed 12 May. 2022.

[14] Zaidi, A, and M Oussalah. Forecasting Weekly Crude Oil Using Twitter Sentiment of US Foreign Policy and Oil Companies Data. IEEE, 2018, pp. 1-8. https://doi.org/10.1109/IRI.2018.00037.

[15] Zhao, Lu-Tao et al. "Forecasting Oil Price Using Web-Based Sentiment Analysis." Energies, vol 12, no. 22, Nov. 2019, pp. 4291-4309. https://doi.org/10.3390/en12224291.