ISyE 6416 - Spring 2022
Project Report

Is the public sentiment on crude oil price related to its underlying hidden state in HMM?

Submission by

Ansh Kuhikar
Rishi Dasgupta
Vamsi Krishna Chakravarthy Ravilla

Abstract / Executive Summary

In the modern world, social media and its ever-increasing impact on the societal perception of any asset has proved to be a significant mover for the world's financial markets. Among all the assets, crude oil impacts every level of the economy - from production to supply chain, and in turn effects the purchasing power of the customer.

The variation of crude oil price is a sequential data and is routinely modeled through Hidden Markov Model (HMM). However, an explicit understanding of the factors that directly determine the hidden states remains unknown. In this study, we explore if the public sentiment of an asset (crude oil) explains the hidden states of HMM and serve as an explicit proxy to the hidden states. For this study, the relationship between oil prices modeled as HMM and public sentiment on crude oil is investigated over the past 8 month period ($\sim$ 150 trading days).

For crude oil prices, West Texas Intermediate (WTI) basket of prices listed on NASDAQ is considered (ticker - CL:NMX) and to capture the public sentiment on crude oil prices, Twitter news headlines from Bloomberg energy, FT Energy, Energy Live News, Financial Times and New York Times are analyzed using FinBERT. Based on this analysis, only a weak correlation is identified between the hidden and sentiment states over the time period in consideration.

1 Problem Statement

Crude oil is an indispensable energy source and the driver of any economy. Besides, crude oil is the backbone of global commerce and trade - it affects everything from flight ticket prices to everyday groceries. Often times these prices are a reflection of the supply and demand for crude oil as well as the overall economy.

Ever since the onset of pandemic, and more recently the Ukrainian crisis, this supply and demand has fluctuated fiercely. This resulted in negative crude oil prices (Fig. 1) during the lockdown [1] and soaring prices (Fig. 2) more recently due to the Ukraine conflict [3, 4]. All these price fluctuations are a result of news about some crisis and policies related to that crisis. Hence, we hypothesize that news articles and their sentiment are driving factors of oil price fluctuations.

In this study, we set out to test this hypothesis by constructing a Hidden Markov Model (HMM - a popular sequential data model) to determine the underlying state of the oil price (ex. bull, bear, consolidation). We will then compute a sentiment metric for crude oil based on the news article headlines from top media agencies' Twitter handles. The final step would be to look for a statistically significant correlation between the sentiment during time of crisis and the behaviour of the oil commodity.

Section 2 details the data used in this investigation and the sources through which it is obtained. The HMM and sentiment analysis methodology is discussed in section 3 and the findings of the study are summarized in section 4. Section 4 also details the potential drawbacks of this analysis and provides directions for future work.

2 Data Sources

2.1 HMM data

There are two main types of crude benchmarks: West Texas Intermediate (WTI) and Brent crude oil. The former is traded on New York Stock Exchange (NYSE), and is centered around US, and North American (Canada and Mexico) oil production. The Brent crude is based on the oil produced from the middle East/OPEC and Russian urals basket [2]. Even though the geographic

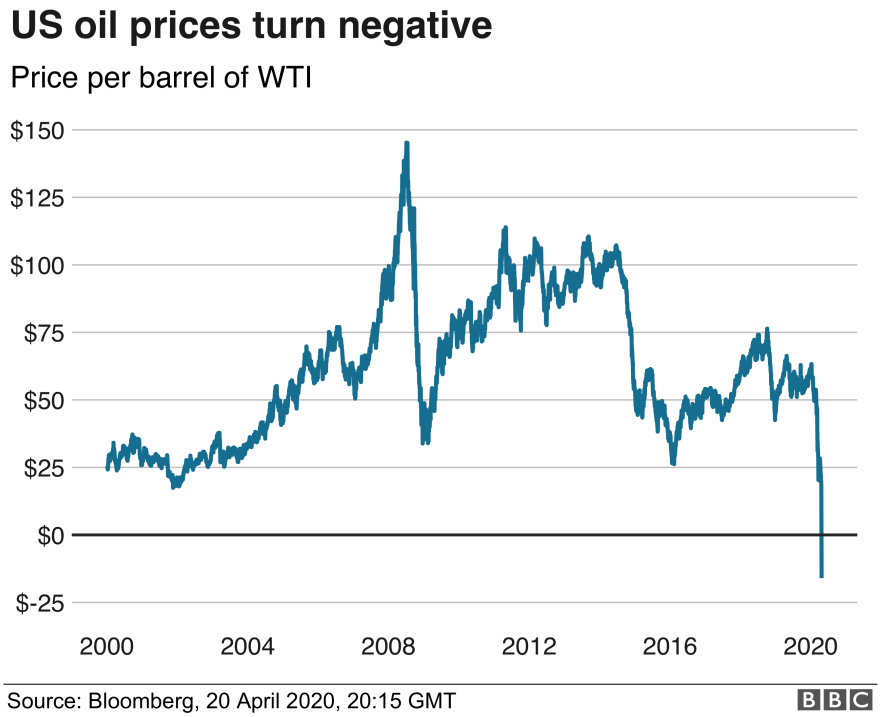

Figure 1: Oil prices drop below zero due to lack of demand during the pandemic.

locations are different, given the global market for types of crude oil, both WTI and Brent oil prices are closely linked to each other. In this paper we focus on WTI because it is the most well known ticker on all major American stock/commodity exchanges (ex. NYSE, NASDAQ) and being US based, it will be more affected by western news media.

To construct the HMM model, we analyze WTI crude oil ticker (CL:NMX) on NASDAQ, which follows crude oil prices. This data is available on Yahoo Finance under the ticker CL:NMX which can be scrapped by the *yfinance* package to obtain ticker, open, high, low, close prices for the stock. The close price is used as the observation to construct HMM. Further details are discussed in the Methodology (Sec. 3). About 5 years of data was obtained towards HMM models.

We chose to analyze this ticker on stock market rather than the spot price on commodities market because the commodities markets is open almost round the clock. This makes it very difficult to ensure temporal causality. On the contrary, the WTI ticker trades only between 9:30 AM and 4:00 PM during the weekdays. Thus, we can easily define Open-High-Low-Close prices of the asset.

2.2 Sentiment Analysis data

In order to capture the sentiment for oil prices, we turn to Twitter feeds of major financial news agencies. Specifically their energy sections, as they will have the most relevant information. Additionally, these are open-access data, and not access-controlled articles that could be targeted towards professionals. It is the open data that forms the bulk of public perception. Thus, by analyzing open-access data, we aim to capture the broader public perception and not just the niche group of professionals in the energy sector.

The news agency handles we chose are: *Bloomberg Energy*, *FT Energy*, *Energy Live News*, *Financial Times* and *NY Times*. Including more might lead to repeated articles and sentiments.

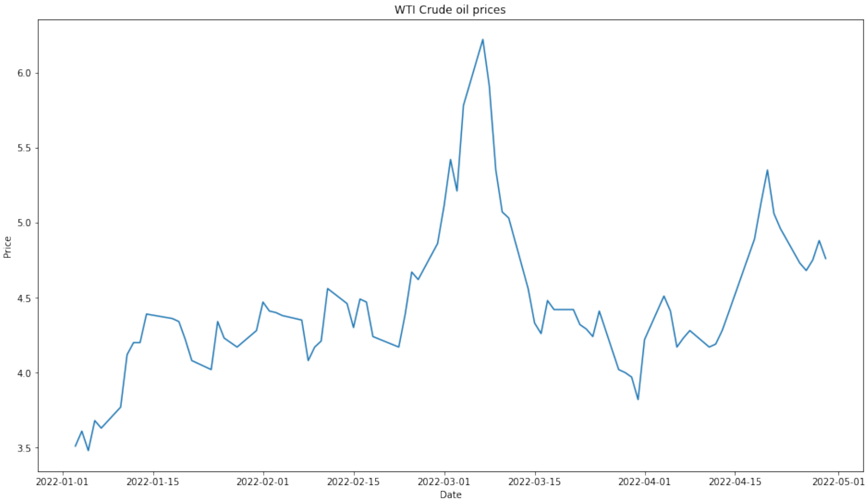

Figure 2: Oil prices shoot up due to Ukraine crisis of 2022.

One could argue that if additional articles/handles are considered they would provide a more accurate assessment on oil prices. For instance, the broader stock market index is strongly effected by energy market and vice-verse (ex. poor energy demand typically indicates a slowdown and this is seen as fall in stock index). However, for this proof of concept, only energy related articles are considered, and broader data sources to capture sentiment may be pursued as a future study.

We scraped Twitter feeds of the above mentioned Twitter handles for the day prior to the close data considered. Stated differently, the sentiment on the crude price for Tuesday is captured by the Twitter feed from Monday. This ensures that we do not look ahead in future and maintain causality (i.e., existing sentiment affects future price and not the converse). Using Twitter *developer* access, we employ the Twitter scrapping package *tweepy* [5] which returns user tweets in a given date range. These tweets are then sorted on a daily basis. The FinBERT package is used to evaluate the underlying sentiment on crude oil using these tweets (further discussion in Sec. 3). As alluded to in the above section, one needs to ensure that we do not look ahead in the future while analyzing the current sentiment. Therefore, to predict the sentiment of the current day, we only consider the tweets posted on the previous day.

3 Methodology

The objective of this investigation is to determine whether the public perception/sentiment of an asset (in this study, WTI crude oil) correlates with the hidden states of the HMM that model the same asset. Therefore, there are two aspects to this study. A schematic describing the procedure is shown in Figure 3.

3.1 Hidden Markov Model (HMM)

The price WTI Crude oil, like any other asset (ex. stock, gold), moves in phases - bull, bear, consolidation, indecisive, etc. The distribution of the price of an asset during the bull phase is likely to be very different from those corresponding to the consolidation or bear phases and vice-versa. However, one cannot observe the asset's current state - instead, we can only predict/estimate the likely state based on the observed price changes. Therefore, we model the daily close price of crude

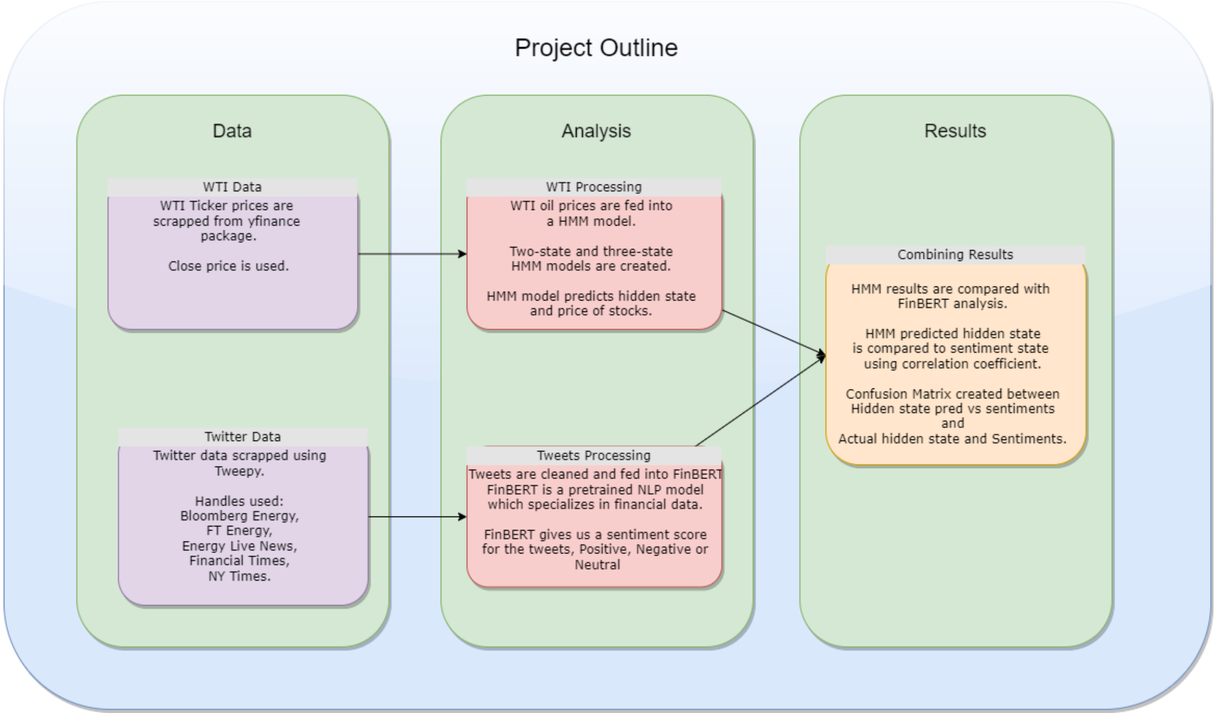
Project Outline
Data
Analysis
Results
WTI Data
WTI Ticker prices are scrapped from yfinance package.
Close price is used.
Twitter Data
Twitter data scrapped using Tweepy.
Handles used: Bloomberg Energy, FT Energy, Energy Live News, Financial Times, NY Times.
WTI Processing
WTI oil prices are fed into a HMM model.
Two-state and three-state HMM models are created.
HMM model predicts hidden state and price of stocks.
Tweets Processing
Tweets are cleaned and fed into FinBERT FinBERT is a pretrained NLP model which specializes in financial data.
FinBERT gives us a sentiment score for the tweets, Positive, Negative or Neutral
Combining Results
HMM results are compared with FinBERT analysis.
HMM predicted hidden state is compared to sentiment state using correlation coefficient.
Confusion Matrix created between Hidden state pred vs sentiments and Actual hidden state and Sentiments.

Figure 3: Workflow: Firstly, the the necessary data is obtained - stock price for HMM and tweets database for sentiment analysis. Further, through these analyses, the underlying hidden states and sentiment for the next day are predicted. Finally, a comparison between the predictions of the two methods is then made.

oil as a first order Hidden Markov Model. A schematic of the process is shown below in Figure 4. The time horizon of the analysis T is set to 30 and the emission matrix is assumed to be a univariate Gaussian distribution $\phi_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ - one for each hidden state. In the current analysis, 2 state and 3 state HMM models are considered.

In general, the number of hidden states required to model the data is an unknown hyperparameter. Previous studies [6] have shown that 2-3 hidden states model the stock prices accurately. There is a strong likelihood that the observation could hold good for the current analysis as well. One could follow a similar approach and determine the optimal number of hidden states for the HMM using model criteria such as likelihood, BIC, AIC. However, such an approach was not pursued because the objective of this study is to correlate the hidden states to the underlying sentiment on crude oil, and not to obtain the most accurate predictive model for crude oil prices. The sentiment analysis from FinBERT returns a *softmax* label for 3 categories: positive, negative and neutral. Therefore, only 2 hidden states (positive - negative) or 3 hidden states (positive - negative - neutral) can be considered.

Another important hyperparameter in HMM model is the time frame T of the observation sequence to be modeled. We consider past T observations to model the HMM and make predictions at $(T + 1)$. In effect, we have a rolling window for observations for training and test data. This is illustrated in Figure 5.

Note that for a 3 state HMM, 14 parameters are to be estimated (2 for prior, 6 for emission matrix, 6 for transition matrix), while for 2 state HMM, 7 parameters are to be estimated. With this in mind, three choices for T are explored in this study: $T = 20, 30, 50$. Due to the limited training data available for $T = 20$, convergence issues were encountered. While no such issues were

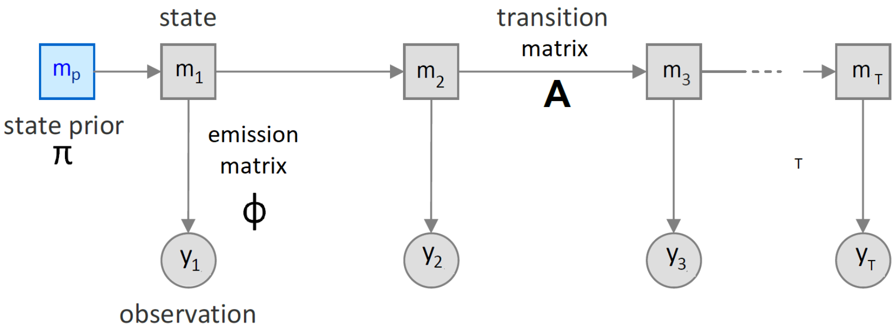

Figure 4: Schematic of the Hidden Markov Model. Using the observation y_i, the parameters of the HMM (transition matrix A and parameters of the emission matrix ϕ) are determined using Baum-Welch algorithm. Once these parameters are determined, the most likely observation at $T+1$ and the underlying hidden state is predicted.

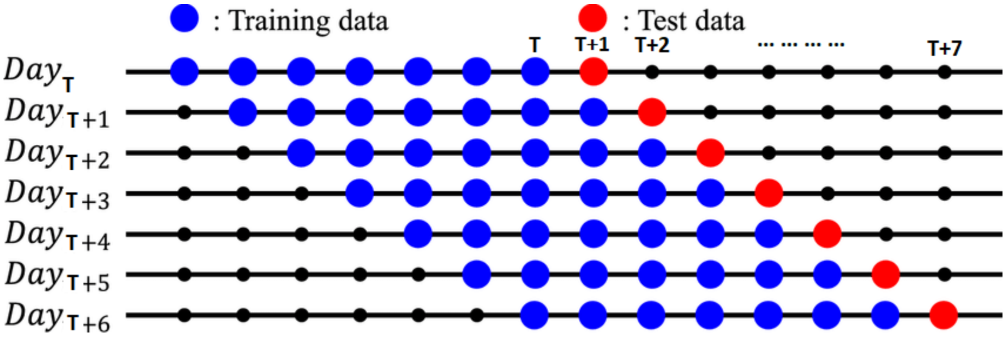

Figure 5: Illustration of sliding window used to train HMM and predict next day price.

observed for $T = 30$ and $T = 50$. Also, the performance 2-state and 3-state HMMs were quite similar for $T = 30$ and $T = 50$. More than 98% correlation (over a 5 year period) was observed between the true and predicted values. Therefore, $T = 30$ is chosen throughout this analysis.

Similarly, to predict the evolution of crude oil prices, one could choose other metrics as observations of the HMM instead of the price y_i. Popular choices for observation variable to model the price evolution include: $(y_{i+1} - y_i)/y_i$ and $log(y_{i+1}/y_i)$. Unlike other studies, where the emphasis is model the final price of crude oil, our objective is to accurately model the crude oil price, observations of HMM (if different from price) and the underlying hidden state. Therefore, the close price y_i is chosen as the observation in the HMM model because of the superior performance in accurately predicting the hidden state and observation for the test data point.

3.2 Inferring hidden state from emission matrix

Using moving/rolling window of data (shown in Fig. 5), the parameters of the HMM are estimated via Baum-Welch algorithm. These computations would be performed using *hmmlearn* package in Python. Thus, we obtain prior and transition matrices along with the mean and variance of the emission probability for each hidden state. However, the sequence of the hidden states is arbitrary - for one instance of moving window, the hidden state 1 could have the highest mean while in the next time step, it could become hidden state 3. In fact, for the same time step, a different random initialization could result in a different ordering of hidden states. Therefore, the

HMM model	Hidden states correlation	Observations/price correlation
2-state HMM	93.2%	97.8%
3-state HMM	89.7%	98.3%

Table 1: Correlations between the predicted and actual hidden states, and predicted price and actual price (observation) of crude oil for 3-state and 2-state HMM.

hidden states are sorted by the mean of their corresponding emission probability - this ensures that the interpretation stays consistent across time steps and is interpretable.

In this study, the hidden states are sorted by their means in ascending order. Hence, for the 2-state HMM, the hidden state 0 can be interpreted to negatively impact (decrease) the price of crude oil, while the hidden state 1 indicates a positive impact (increase) on crude oil price. Similarly, for a 3-state HMM, the hidden states 0 and 2 denote the negative and positive effects on crude oil price, whereas the hidden state 1 denotes neutral impact on crude oil price. This interpretation can be further illustrated by tracking the variation of means of the emission function for each hidden state. Figures 7 and 6 show the variation of means corresponding to each hidden state for the last 300 timesteps (this roughly equals 1 year of trading period - exchanges are remains closed on weekends and national holidays). For both models, we see that a close match is observed between the predicted price and actual price (Truth) of crude oil (> 98% correlation). The results are tabulated in Table 1. Furthermore, the regions of price raise and fall in prices match with the qualitative behaviour of the corresponding hidden states.

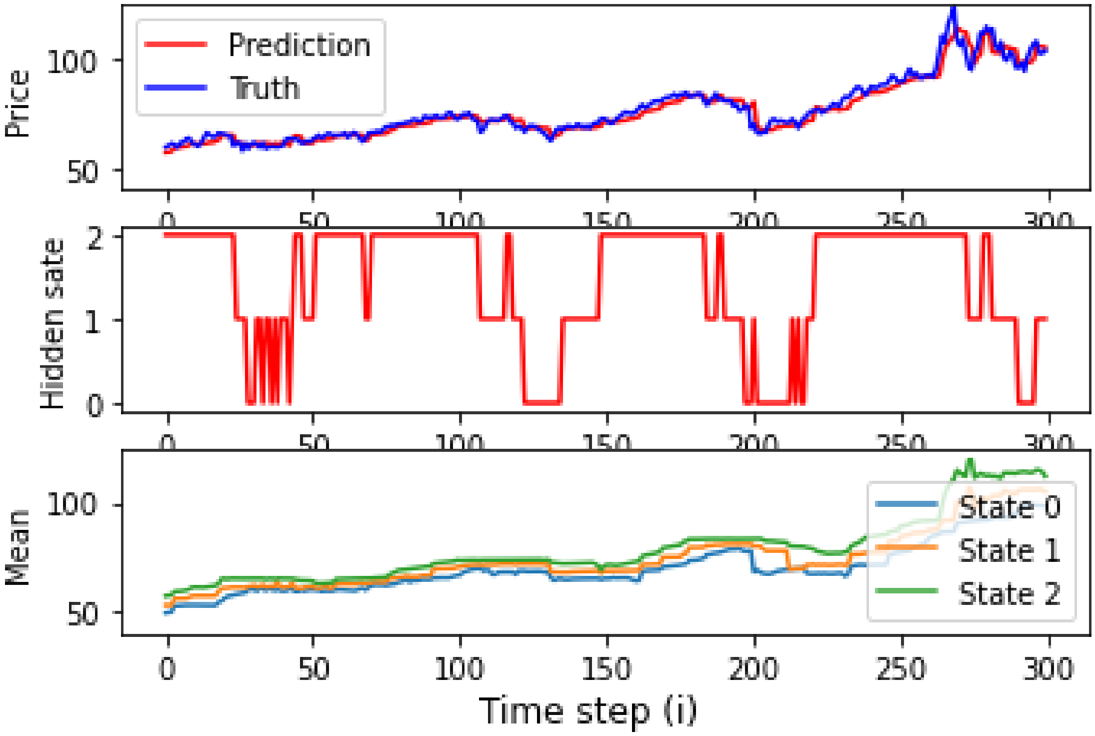

Figure 6: Evolution of observations (price), hidden states and their underlying mean of emission function for the 3 state Hidden Markov model.

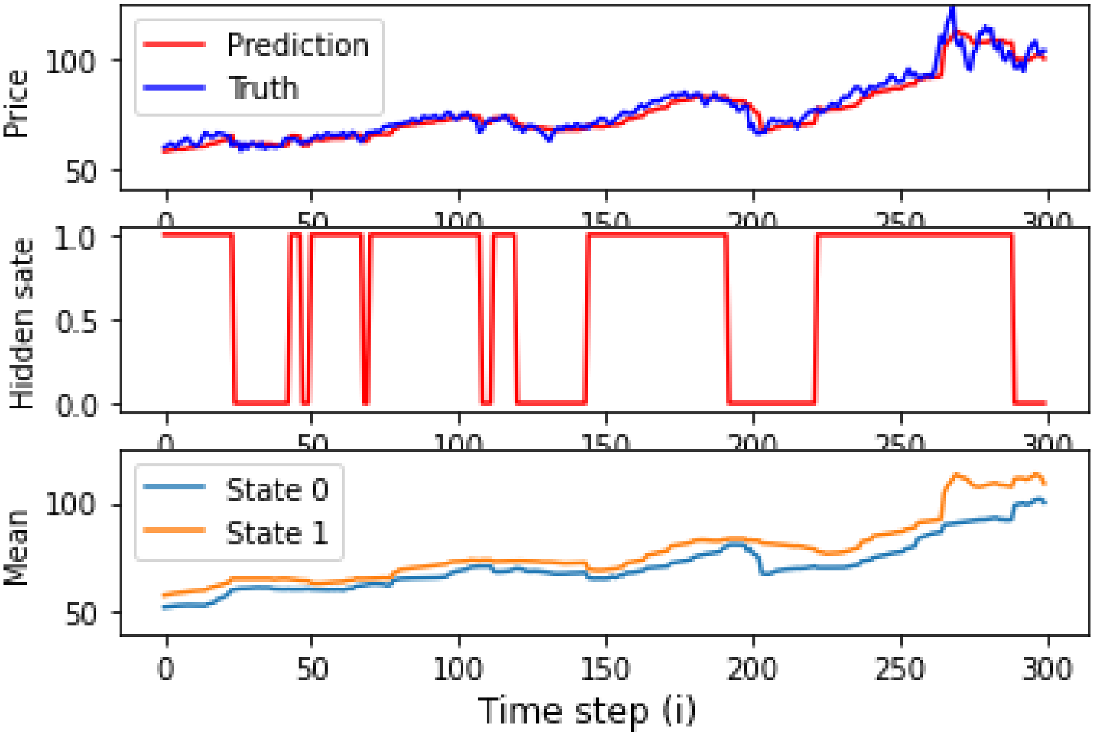

Figure 7: Evolution of observations (price), hidden states and their underlying mean of emission function for the 2 state Hidden Markov model.

In summary, the HMM models with univariate Gaussian emission accurately predict the observations as well as the hidden states. The interpretation of the hidden states qualitatively matches the expected impact on price.

3.3 Predicting future observation and hidden state through HMM

Once the HMM parameters are estimated through Baum-Welch algorithm, the prediction at $T+1$ is obtained such that

$$y_{T+1} = argmax_{y_{t+1}} \mathcal{P}(y_{T+1}|y_1, y_2, ..., y_T, \pi, A, \phi),$$

where y_i is the close price of the crude oil at time i, and (π, A, ϕ) are parameters of the HMM. A grid of values is assumed for y_{T+1}, and each value on this grid is tried as guess value for y_{T+1} and the value that maximizes the likelihood is chosen at the observation a next time step. Care is taken to ensure that the range of values assumed for y_{T+1} is wide enough such that the maxima occurs within the assumed grid of values (about 50 guess values of y_{T+1} are tried for each time step).

After the future observation y_{T+1} is predicted, the underlying hidden state is obtained as

$$(m_1, m_2, ..., m_{T+1}) = argmax_{m_1, m_2, ..., m_{T+1}} \mathcal{P}(y_1, y_2, ..., y_{T+1}, m_1, m_2, ..., m_{T+1}|\pi, A, \phi),$$

where m_i is the hidden state at time i. This decoding problem is solved via Viterbi algorithm available in *hmmlearn* package.

In the next section, we explore the sentiment analysis of crude oil price - if a strong correlation between hidden states and sentiment exists, it could serve as an observable proxy for hidden states and potentially improve price predictions.

3.4 Sentiment Analysis - FinBERT

Gauging the sentiment of a text statement is an inherently complex task. Over the years, various methods based on bag-of-words approach, such as GloVe, and LDA were used for sentiment analysis [7] with limited accuracy. In recent years, deep neural network architectures progressed significantly in sentiment analysis and outperformed traditional NLP methods. In fact, deep models trained on massive datasets such as BERT [8] and ELMo have even proved to be a good basis for *transfer learning* - where most of the neural network layers are frozen and only the final layer(s) are re-trained to suit the application at hand.

In this study, FinBERT is used - one such pre-trained transfer learning model. FinBERT [9] is used to determine the sentiment of the underlying text (in this case, tweet). As the name suggests, FinBERT is a version of BERT that is tuned using financial data - 48,000+ financial news articles published in Reuters between 2008 and 2010, and Financial PhraseBank corpus of annotated data were used to fine tune BERT for sentiment analysis of text with financial jargon.

In brief, BERT (Bi-directional Encoder Representations from Transformers) is trained to predict the missing/masked word in the sentence in both directions (i.e, the sentence is input in forward as well as reverse direction). As we can see, in its native form, BERT is not trained to predict sentiment of the text. However, given the deep neural network of the model, it grasped the underlying semantic meaning. Therefore, the last layer of the BERT is replaced by *tanh* or logistic layer and this modified layer is re-trained [10]. This is called transfer learning - a popular and currently the go-to approach for many Natural Language Processing (NLP) and Computer Vision tasks. This is particularly powerful when the data set available for training is small.

Given that FinBERT is already tuned for financial reports of reputed news agency and financial institutions, the distribution of the text we are collected from their Twitter posts is likely to have similar distribution. Hence, we directly use FinBERT classifier as a pre-trained model without any additional training.

Overall, once the hidden states from HMM and the sentiment of the the Crude Oil based on Twitter feed are obtained, we will compare the two metrics to determine their relationship through metrics such as confusion matrix and correlation coefficient.

4 Evaluation and Result

4.1 Results

We can infer that both the 2-state and 3-state Hidden-Markov models predict oil price fluctuations with around 95% accuracy over the time period in consideration. There is however, a delay or lag in oil price prediction in the range of 1-2 days, which is caused due to the fact that the Viterbi algorithm does not switch states for mild deviations. This results in a delay in switching between the hidden states and consequently the price prediction.

On the correlation between hidden states and public sentiment, a relatively weak correlation is observed: a 10.65% correlation for 2-state HMM and a 4.17% correlation for 3-state HMM.

Correlations	**2-state HMM**	**3-state HMM**
Hidden state (HS) correlation (predicted)	95.18%	89.66%
Observation/Price change correlation	95.13%	96.24%
Sentiment State (SS) correlation (predicted)	10.65%	4.17%
Sentiment State (SS) correlation (actual)	8.96%	6.43%
Confusion Matrix (HS/SS predictions)	$\begin{bmatrix} 36 & 16 \\ 58 & 50 \end{bmatrix}$	$\begin{bmatrix} 2 & 21 & 1 \\ 2 & 38 & 0 \\ 14 & 71 & 5 \end{bmatrix}$
Confusion Matrix (HS/SS actual)	$\begin{bmatrix} 30 & 17 \\ 58 & 49 \end{bmatrix}$	$\begin{bmatrix} 1 & 27 & 0 \\ 4 & 28 & 1 \\ 13 & 75 & 5 \end{bmatrix}$

Table 2: Results Summary

4.2 Conclusion

The Hidden Markov models are found to accurately predict the fluctuations in oil prices over the 150 days time period under consideration. A small lag in price prediction in the range of 1-2 days has been observed which is consistent with similar studies performed. Augmenting the hidden state with external information (ex. Input-Output HMM [6]) could help address this lag. As accurate price was not the primary objective of the study, this path was not explored.

Based on the results obtained in Table 2, we infer that the sentiment states obtained from the sentiment analysis have weak correlation with the hidden states obtained from both the 2-state and 3-state HMMs. Therefore, we conclude that the sentiment analysis alone does not serve as a reliable proxy for oil price fluctuations.

4.3 Future study

In the future, we may attempt to identify and address possible causes for a lack of match between the hidden states and the sentiment states as follows:

1. Sentiment is seen to be an unreliable predictor for daily price variations in this investigation. Only big shocks/moves are captured by the public sentiment. This suggest that the public sentiment quickly dissipates and thus stays irrelevant for most of the time. Therefore, modeling the daily Twitter sentiment (say via higher-order Markov model, moving averages) to capture the past sentiment could make sentiment analysis more relevant to capture daily price variations.

2. The time period under consideration could be expanded to cover periods of lower or more consistent volatility. A Granger Causality test can be conducted to identify whether a significant correlation exists over longer time periods.

3. Limited Data is considered which implies that the sentiment is not accurately captured. Instead of Twitter feeds, detailed news articles and opinion pieces may be considered which may be more representative of the overall sentiment.

4. Utilizing sentiment analysis in conjunction with other fundamental and technical indicators could better predict and provide an intuition for the hidden Markov states.

References

[1] https://www.bbc.com/news/business-52350082

[2] Historical Dictionary of the Petroleum Industry. Lanham, MD: Scarecrow Press. ISBN 0-8108-5993-9

[3] https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/08/fact-sheet-united-states-bans-imports-of-russian-oil-liquefied-natural-gas-and-coal

[4] https://www.nytimes.com/2022/03/31/business/energy-environment/biden-oil-strategic-petroleum-reserve.html

[5] Roesslein, J., 2020. Tweepy: Twitter for Python! URL: https://github.com/tweepy/tweepy

[6] Christensen H., Godsill S., & Turner R. E. (2020). "Hidden Markov models applied to intraday momentum trading with side information", arXiv preprint arXiv:2006.08307.

[7] Nguyen T. H., & Shirai K. (2015). "Topic modeling based sentiment analysis on social media for stock market prediction", Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1354-1364).

[8] Devlin J., Chang M.W., Lee K. & Toutanova K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805.

[9] Araci D. (2019). "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models", ArXiv, arXiv:1908.10063.

[10] Sousa M. G., Sakiyama K., de Souza Rodrigues L., Moraes P. H., Fernandes E. R., & Matsubara E. T. (2019). "BERT for stock market sentiment analysis", In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1597-1601). IEEE.