Using Sentiment Analysis on 10-k Filings to Predict Oil Prices

Peter Kowalchuk

School of Professional Studies, City University of New York

Data 698: Analytics Master's Research Project

Jamiel Sheikh

November 14, 2022

Introduction

Predicting financial and economic indicators, including equity and commodity prices, is a significant data analysis use case. Natural Language Processing (NLP) techniques such as Sentiment Analysis are finding their way into Financial Analysis. A review of the established literature reveals that Sentiment Analysis of company filings and earnings reports produces meaningful equity investment insights. However, predicting the price of a commodity with Sentiment Analysis has not been widely studied.

This study attempts to prove that it is possible to predict the price of a commodity from the content of the Security and Exchange Commission (SEC) 10-K filings. These reports must be filed annually by all publicly traded companies. The overall objective is to determine if these reports contain enough information about the broader market that make it possible to predict the price of a commodity. Oil was selected as the commodity to analyse mainly because of its significant economic impact. The study divided reports based on the type of oil company making the filling and included a control group of non-oil sector companies to account for potential bias.

Sentiment was calculated for reports of companies included in the study, with data quality and statistical significance considered to reduce noise and company and industry sector bias as much as possible. From the Sentiment data generated, models were constructed to predict oil prices. The resulting models were of different quality levels depending on the data generated. Both low-quality and fair-quality models were developed.

While the study did not conclusively confirm the stated hypothesis, it did conclude that the possibility does exist of generating a model to predict the price of Oil. The study concludes with suggestions for future work that may make it possible to confirm that 10-k filings do contain broad market information.

Literature Review

There is rich literature on prior work done on Sentiment Analysis performed on U.S. SEC (Security and Exchange Commission) company filings. G. H. Soong and C. C. Tan (2021) found that it is possible to accurately determine the Sentiment of a company filling using text analysis tools. They calculated the performance of different models used to assess the Sentiment of the filings against manual Sentiment labelled reports. While the models they tested produced different accuracy levels, they proved that it is possible to correctly determine a filings Sentiment with an accuracy of up to 90.08%. In his work, Pablo (2009) went so far as to state that Sentiment Analysis can produce results comparable to humans.

Gao, Kampas, and Rinne (2018) found that researchers have mainly worked on Sentiment from survey opinion polls, newspaper articles, newswires, and information from social networks, microblog websites, and internet message boards such as Twitter. Their literature review did not highlight the use of SEC reports. Researchers Gupta and Chen (2020) have been able to prove that Sentiment Analysis can be used to enhance the prediction of a company's stock price. In their work, they used Sentiment Analysis derived from StockTwits to improve their stock price change prediction by an average of 1.7%. Other researchers have done similar work using other text for their Sentiment Analysis. Seals and Price (2020) improved stock price forecasting using Sentiment from New York Times articles.

A significant amount of previous literature shows the potential of Sentiment Analysis in financial applications and future predictions. "Laws and regulations prohibit companies from making material or misleading statements" (U.S. Securities and Exchange Commission, "How to Read a 10-K," 2011) in SEC 10-K reports. This makes them a good candidate for Sentiment Analysis research. The SEC provides rules that companies must follow to complete these reports, including the information that must be contained in each section of the report. Researchers Azmi Shabestari, and Romero (2022) have investigated the Sentiment and tone of the language used in some of the report's sections. In their work, they found that some sections contain significant Sentiment and tone that can be used in further analysis. Bao and Datta (2014) successfully used 10-K text data to analyse investors' perception of risk, showing us that there is indeed potential for these reports to be a source of valuable financial analysis beyond the reporting companies.

While there is a significant amount of work on Sentiment Analysis, stock prediction applications, and work with SEC reports, there is much less research using these reports and these techniques to predict broader financial indicators. The SEC guidelines on 10-K reports state that companies can use different sections of the reports to comment on the competitive forces in the market, seasonal factors, entire economy, geography or industry sector risks, and exposure to market and commodity price risk. This suggests that the reports can be used to predict the reporting company's stock performance and broader financial and economic indicators, assets, or others. Rolnicki. (2018), for example, looked at measuring the risk in the financial sector using 10-K reports, but the study did not include the prediction of a specific asset. Therefore, there is a research gap pertaining to using 10-K reports to predict the price of a specific asset.

This study attempts to fill the identified gap by analysing multiple reports from different kinds of companies using Sentiment Analysis and using the results to find a prediction indicator of oil prices. By leveraging the information from the SEC guidelines, the study asks if it is possible to use this data source to predict oil prices. Since laws and regulations limit the presence of false or misleading statements in the SEC 10-K reports, broader market information contained in them could represent a reliable source.

This study will include a thorough analysis of the proper content of the report since selecting different content will impact the Sentiment Analysis results. The study will also produce results for different groups of companies to discover a set of companies that can accurately predict oil prices from 10-k reports' Sentiment data. Researchers such as Feldman et al. (2009) have shown that changes in the tone managers use in 10-K reports can greatly influence any analysis in which they are used. General issues with using sentiment analysis in financial applications should also be considered, as highlighted by the work of researchers such as Xing et al. (2020). To avoid these issues, the selection of the correct data to use is a critical aspect of the study.

Hypothesis

The main goal of this study is to answer the question: Do SEC 10-K filings have enough broad market information to predict a commodity's price? Hypotheses were formulated to answer this question. The study will use SEC 10-K reports and the Sentiment Analysis NLP technique to prove or reject these hypotheses. The question posed for the study will only be answered negatively if all hypotheses are rejected.

Hypothesis 1: it is possible to design a linear regression prediction model that, with reasonable accuracy, can predict the price of oil from SEC 10-K report filings' computed Sentiment Data.

Hypothesis 1 will be rejected if it is concluded that it is not possible to build a model with what can be considered a good Mean Squared Error (MSE). A good threshold of $2 for MSE is used in this study based on prior results from other researchers such as Tebyanian and Hedayati (2014).

The study will also consider that while defining a predictor model is not possible, the oil price variance can be explained by the Sentiment in the reports. This would also indicate that reports do contain broad market information.

Hypothesis 2: it is possible to design a regression model that explains oil prices using Sentiment data from SEC 10-K reports as the predictor variable.

Any variance in oil price explained by the model would indicate that the reports contain broader market information. An R-Square threshold of 0.5 defines enough or substantial, broad market information to keep the study.

If hypothesis 2 is not rejected, the study will analyse models derived from reports from only companies that are directly tied to the production and commercialization of oil.

Hypothesis 3: SEC 10-K reports from companies that operate directly in the oil sector produce better Sentiment based regression models for oil prices than companies that do not directly operate in the oil sector.

Finally, the study will also analyse the possibility of a sub-set of oil sector companies being a better predictor of oil prices.

Hypothesis 4: There exist a group of oil sector companies that can better produce regression model oil prices from their SEC 10-K filings.

Methodology

This study was conducted using data from 10-K report filings from 2000 to 2021 for thirty-nine companies sourced from the Security and Exchange Commission's EDGAR site. Oil price data was sourced from the U.S. Energy Information Administration's web site for the same period. Both Brent and West Texas Intermediate prices were sourced. Companies in the study were classified into three groups. The first two groups comprise companies directly involved in oil extraction, production, and commercialization. In contrast, the third group consists of companies with no direct ties to the Oil and Gas sector.

The first group included Operator Oil and Gas Companies - labelled Operators. These companies own oil fields and make significant infrastructure investments to extract and bring oil to market. Capital invested by these companies takes a long time to produce returns. This is due to the complexity of planning, designing, constructing, and operating facilities required to bring oil to market. The group brings to the analysis the view of companies with long investment horizons. The group consisted of eleven companies selected to include multinationals that operate across the globe and independents with a regional approach. The group had vertically integrated companies with upstream, midstream, and downstream operations and companies focussed on upstream. Companies operating in conventional and unconventional fields were included. This diversity of company strategies and operating environments capture the view of a wide range of Operators, which helps prevent bias from a particular set of companies.

The second group consisted of Oil and Gas Service Companies - labelled Services. These companies support Operator companies with technology that maximizes oil extraction efficiency. While technology investment might take substantial time to produce returns, in general, these companies can produce returns on capital much more quickly than Operators. These companies bring a short investment horizon point of view to the analysis. A total of eleven Service companies were included in the study. Companies with a local, regional, and global footprint were included, as well as companies with investments in different oil extraction and production technologies. The goal with this wide selection was once again to avoid bias from a particular set of companies.

The third group, Control, consisted of seventeen randomly chosen companies not known to have direct ties to the Oil and Gas sector. Companies in this group were selected to include representation from industries such as technology, retail, automotive, construction, banking, and healthcare. The number of companies that report SEC filings is believed to be around 3,600. This control group is intended to serve as a sample of that population.

A total of 1,208 10-K reports were analysed in this study, including companies in all three defined groups. Microsoft's Cognitive Services Sentiment Analysis service was used for the analysis of all reports. The results from the Microsoft tool were taken at face value. Benchmarking of the selected tool was not within the scope of the study. The chosen tool produced scores for Positive, Negative, and Neutral Sentiments of the text it was presented with. Rather than calculating Sentiment for the entire text contained in the sourced filings, two sections were selected for the analysis. SEC guidelines state that Item 1 in these reports may “include information about recent events, competition the company faces, regulations that apply to it, labour issues, special operating costs, or seasonal factors” (How to Read a 10-K.). Sentiment in this section may then be influenced by the broader market conditions and thus help predict the price of oil. SEC guidelines state that Item 7 is an opportunity for the company's management to discuss business results of the past year, as well as challenges and risks. Similarly to Item 1, this section may produce Sentiment correlated to broader market conditions. Sentiment for these two sections was calculated separately to filter out the noise produced by analysing sections of the report that focus on presenting financials.

Before producing a correlation between sentiment and oil prices, the data was cleaned, and the quality of the sentiment analysis results was inspected. The main goal of data cleansing was to eliminate analysing text with neutral Sentiment. This kind of Sentiment result is interpreted as being produced from text that rather than contain market conditions information, is simply stating company facts and financials. To ameliorate this effect, reports were partitioned into smaller text files. Results with no Negative or Positive content were then eliminated from further analysis. As a second data filtering step, sentiment data for each year were analysed separately to determine the statistical significance of the three sentiment results: positive, negative, and neutral. Data for the period was bootstrapped, and 95% confidence intervals were calculated. For years in which confidence intervals overlapped, sentiment analysis results were not included in the analysis.

The third and last step in the study's methodology was to analyse the correlation between the Sentiment results and the price of oil. Daily oil prices were down sampled to the Sentiment data's frequency by calculating average year prices. From this correlation, a specific Sentiment was selected. The selected Sentiment, Positive, Negative or Neutral, was then used to build a regression model to predict the price of oil. Several regression models were tested. The study's hypothesis was then confirmed or rejected by analysing the results of the regression models.

Study Results

Security and Exchange Commission 10-K filing reports for 39 companies were used in the analysis. Analysis period was from 2000 to 2021. This resulted in a total of 1,208 reports analysed. Oil prices for the same period was collected for both the US West Texas Intermediary and the International Brent benchmarks.

Filings Sentiment data was calculated for Items 1 and 7 of the 10-K reports, to avoid analysing financial heavy sections of the filings. An initial analysis of entire reports sections showed Sentiment analysis results that seem highly influenced by the reporting company and the group they belonged to. For example, in the Services group, one company was found to always show a positive Sentiment, while all its peers showed a clear Negative Sentiment. This can be seen in plot A) of the figure below. The strong Positive Sentiment in the middle of the graph corresponds to a single company, while the Negative sentiment in the left and right are for several different companies. In the Control groups, companies were found to show contradictory Sentiments.

Plot B) below shows two companies showing inverted Neutral and Negative Sentiments.

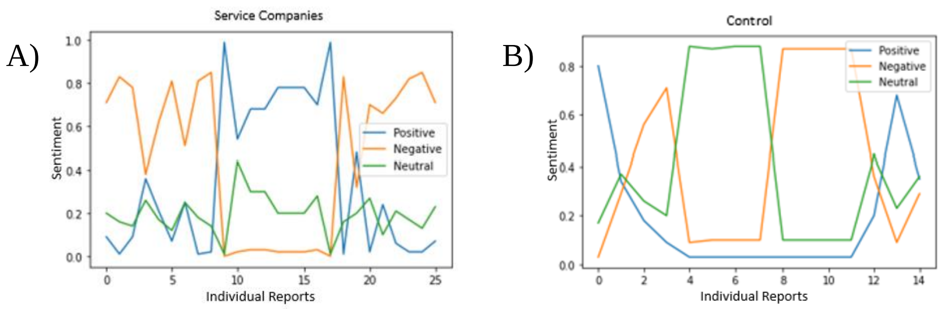

Figure 1. Issues found when analysing entire reports.

The source of these issues could be rooted in the reporting style of each company. To help reduce the effects of this bias, reports were further partitioned into 5120-character sections, thus oversampling the complete dataset. A total of 16,335 sections were produced, each treated independently in the study. Sentiment was calculated for all these sections which provided a more granular analysis of Sentiment in the two report Item sections selected. This finer granularity allows cleaning results that do not contain many Positive or Negative Sentiments. Only 1.34% of the dataset was found to contain poor Sentiment content. This low percentage of poor Sentiment sections provided confidence in the quality of the dataset used. A total of 16,116 sections were available after data cleaning. All sections were labelled with provenance metadata and handled equality in the study.

	Positive	Negative	Neutral	Group	Period	Section	Co	Year
0	0.02	0.00	0.98	Services	P3	1	HAL	2018-01-01
1	0.47	0.40	0.13	Services	P3	1	HAL	2018-01-01
2	0.18	0.59	0.23	Services	P3	1	HAL	2018-01-01
3	0.29	0.57	0.14	Services	P3	1	HAL	2018-01-01
4	0.14	0.57	0.29	Services	P3	1	HAL	2018-01-01
...	...	...	...	...	...	...	...	...
16327	0.13	0.65	0.22	Operators	P3	7	OXY	2021-01-01
16328	0.11	0.80	0.09	Operators	P3	7	OXY	2021-01-01
16329	0.01	0.83	0.16	Operators	P3	7	OXY	2021-01-01
16330	0.00	1.00	0.00	Operators	P3	7	OXY	2021-01-01
16334	0.01	0.00	0.99	Operators	P3	7	XOM	2021-01-01

Table 1. Study dataset sample

It was also essential to understand if the Sentiment data was statistically significant for each year in the analysis period. Statistically significant was defined as yearly Sentiment data that showed distinct Positive, Negative, and Neutral Sentiments. For this to be true, confidence intervals for the annual Means of these three measurements had to be non-overlapping. A first quick view of the entire study period shows how the three Sentiment data with clear distributions are somewhat Normal but clearly overlapping. This plot of the whole dataset, while reassuring of the quality of the calculated Sentiment, cannot be used to assess the statistical significance of all the data. The approach was to amylase the data at the sample rate of the analysis. So rather than looking at the data overall, significance was tested for each year. If a year was found with no statistically significant Sentiment, it would be excluded from the study.

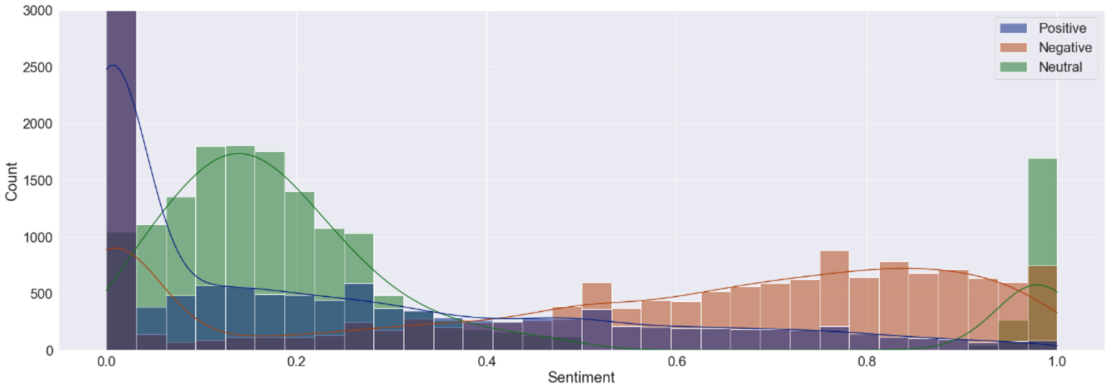

Figure 2. Histogram plot of Sentiment Data in study

To determine Sentiment significance for each year, independently, bootstrapping was performed for Sentiment data for each year. Confidence intervals were calculated for each year and tested for overlap. None was found for the entire study period, including for years where the Sentiment means showed relatively close. The year 2015 serves as a good example, a year in which Positive and Neutral Sentiments were very close. Even for this year, the distribution of the Sentiment means showed very separated.

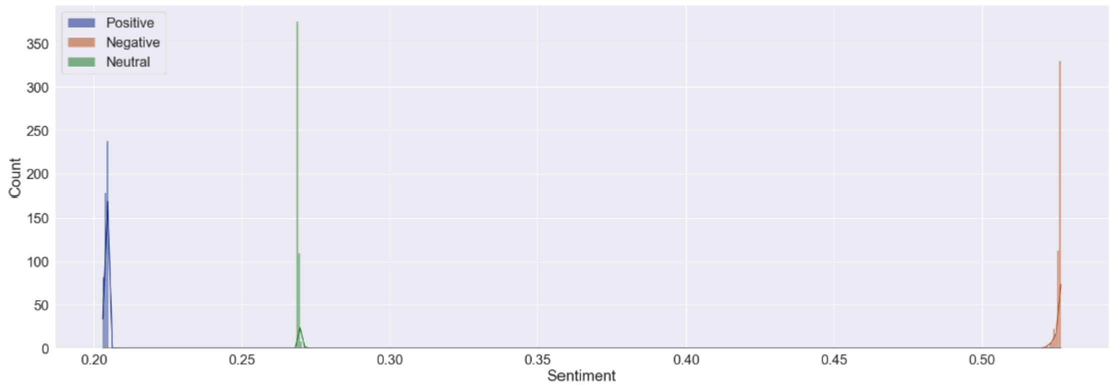

Figure 3. Bootstrap Sentiment data distributions for 2015

Once cleaning and validating Sentiment data was completed, oil prices were merged into the dataset. Oil data was downscaled to the Sentiment data's yearly sample rate. An average of yearlong prices oil prices was used as the year price for the commodity. The study did not include other techniques to down sample prices. With both Sentiment and price data at the same sample rate, it was now possible to plot prices against sentiment and commence correlating both data and determine if it is possible to define a prediction model for the commodity.

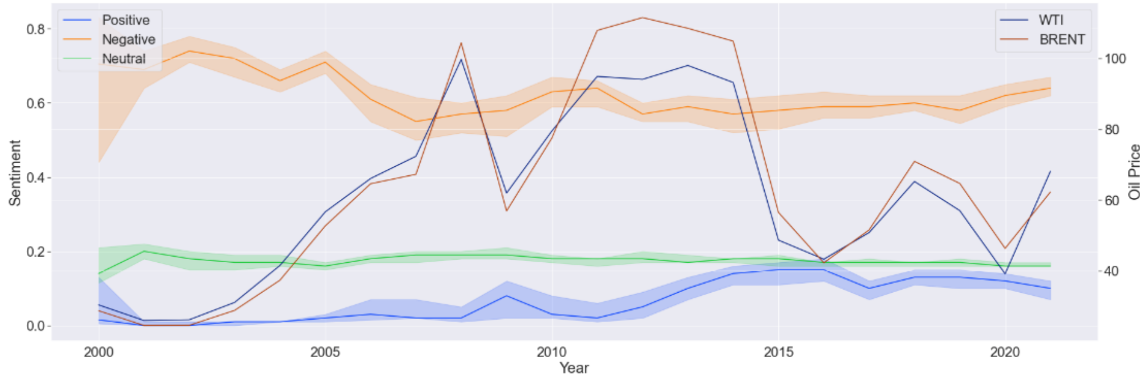

Figure 4. Sentiment and Oil Price dataset plot

To confirm or reject the first hypothesis, the study searched for a model that performed well in making this prediction. Linear models were fitted with the sentiment data as predicting variables and the price of oil as the dependent variable. A quick visual inspection of Positive and Negative Sentiments against WTI oil price shows promising results towards confirming the hypothesis. As seen in the graph below, increasing Positive sentiment correlates with rising oil prices, while decreasing Negative Sentiment correlates with falling oil prices.

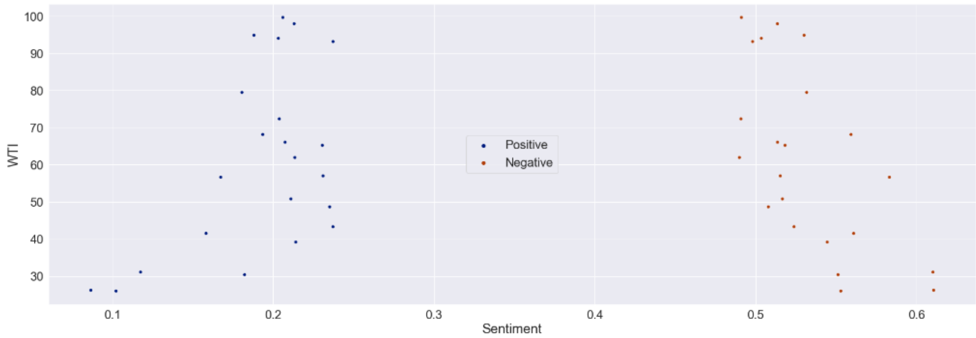

Figure 5. Positive and Negative Sentiment vs. WTI oil price

Because the study's period was of 21 years, or 21 data points, models were bootstrapped to derive a better view of performance. Starting with using data from all first groups of companies, models that included different combinations of sentiment were generated. Model were then analysed and benchmarked against the defined hypothesis rejection thresholds.

The study was unable to find a prediction model with an MSE that met the requirements of the hypothesis. MSE’s for the model calculated were several orders of magnitude greater that the defines threshold of $2. As an example, one model showed a threshold of $192.

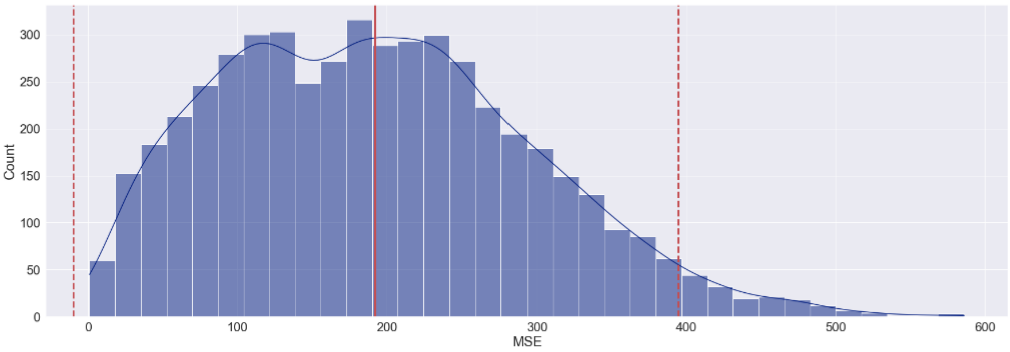

Figure 6. MSE for Linear Model generated with all three Sentiments

Not being able to generate a linear regression model to predict oil prices means hypothesis one is rejected. To work on the other hypothesis, model were analysed for their ability to explain the oil price variance from the reports Sentiment. To do this, the R-squared of the models was inspected and benchmarked against the defined threshold.

While some models produced bootstrapped mean R-Squared that were above the thresholds, their distributions were not Normal, and the resulting confidence intervals made them unsuitable for use.

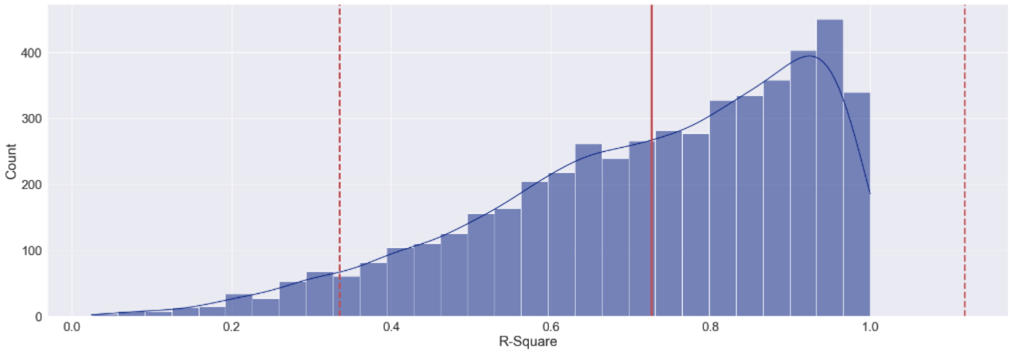

Figure 7. Linear Model generated with all three Sentiments

But Certain combinations of Sentiment data produced better regression models. For example, Positive and Negative Sentiment for all the companies in the study, generated a model with an R-Square metric of 0.585, and a 95% confidence interval from 0.174 to 0.996. While not conclusive, these results left open the possibility that hypothesis 2 could be confirmed.

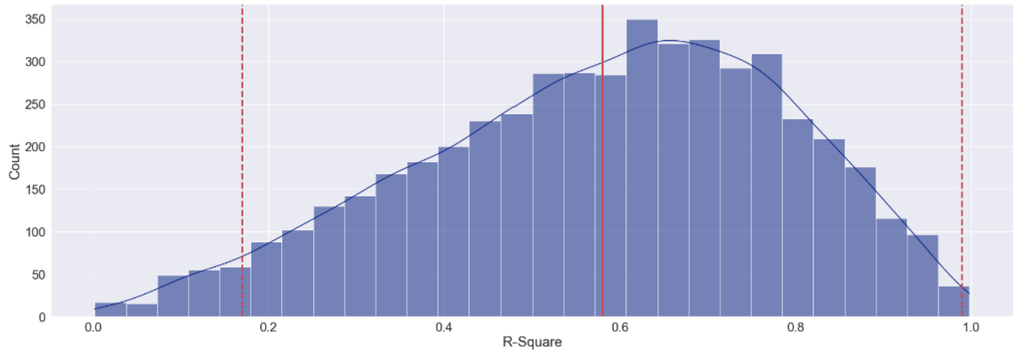

Figure 8. Linear Model generated with Positive and Negative Sentiments

For hypothesis 3, models were generated using data from companies that operate in the oil sector. This group included both Operator and Service companies. While results were not as good as those generated with all companies, R-squares higher than 0.5 were achieved, although confidence interval were still large. As with the previous hypothesis, results left the possibility that a model to confirm hypothesis three might still be achievable.

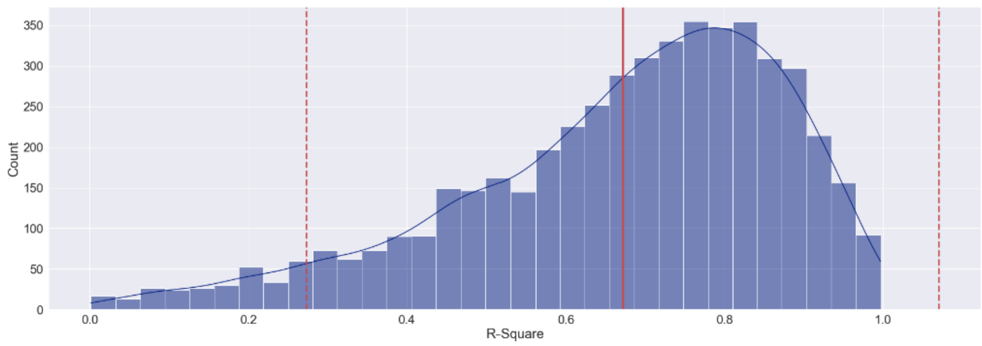

Figure 9. Linear Model generated with Operator and Service Companies

For the last hypothesis, the goal was to determine if either group of Oil & Gas companies were better at predicting oil Prices. It was expected that the Operators group would provide a good amount of insight into the oil market. But results proved otherwise.

The prediction model for Operators, using Negative and Positive Sentiment, which produced the best model, showed a confidence interval from zero to one. These means the model is unable to explain oil prices. Interestingly, the model distribution of R-Square, almost looks binomial, suggesting that the dataset contains two different sub-groups. These maybe a combination of Multinational, National Oil Companies, Conventional or Un-conventional companies.

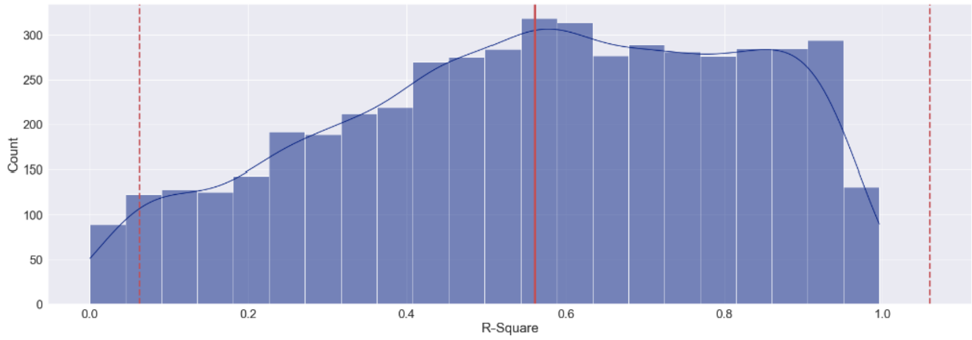

Figure 10. Linear Model generated with Operator companies

Interestingly, model results using Oil & Gas service companies also showed poor results. Similar to Operators, while the R-square mean was a reasonable 0.55, the 95% confidence interval again covers the entire spectrum from zero to one.

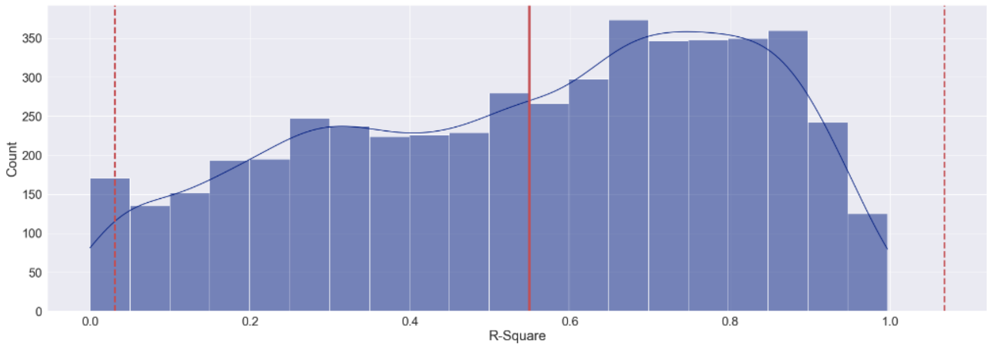

Figure 11. Linear Model generated with Service Companies

Sections of the filings were also analysed separately to understand if different portions of the reports are richer in Sentiment content. Item 1 and 7 of the reports were processed separately. Models generated with Item 1 data were found to be of very poor quality, as ca be seen in the graph below.

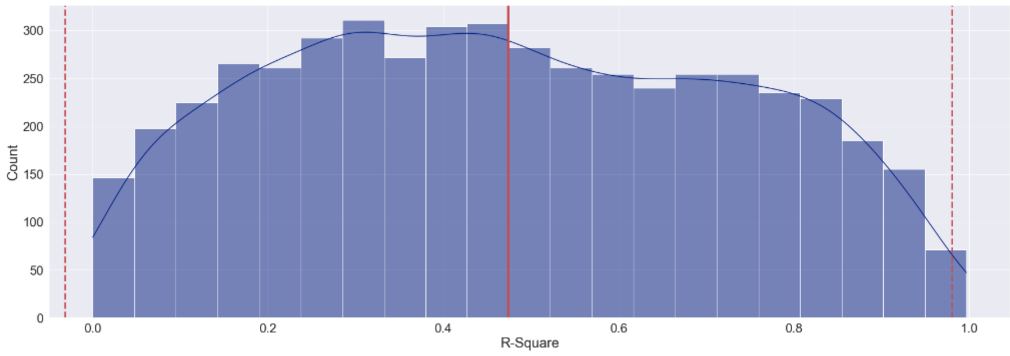

Figure 12. Linear Model generated with all companies using Item 1 only

Models generated with Item 7 Sentiment, produced better results, leaving the option open that models to predict oil prices might be possible using SEC filings Sentiment data.

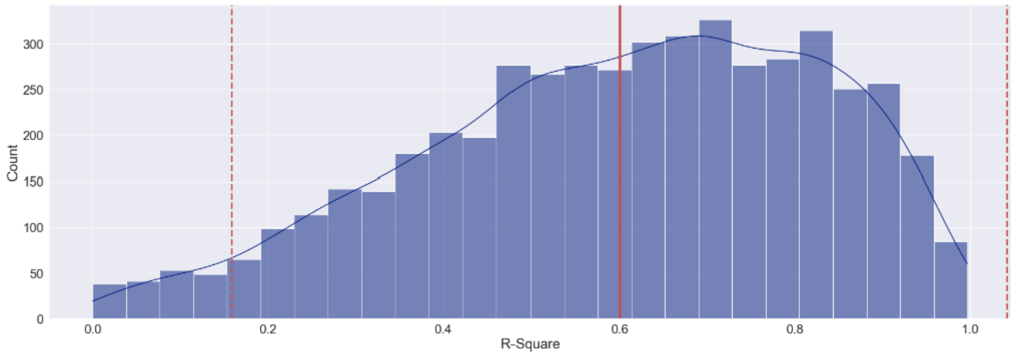

Figure 13. Linear Model generated with all companies using Item 7 only

Conclusions

By computing the Sentiment Analysis of companies' 10-k SEC filings, models were produced in this study that can, up to an extent, explain the price of a commodity. They did not, however, prove that predicting oil prices is possible. Out of the three groups of companies considered, the Control group produced the best results. Although the quality of the models generated using Control group companies was reasonable at best, they do leave the possibility open that a model could be generated using Sentiment data to explain the price of a commodity. The model generated with both Operator and Service companies' Sentiments produced inferior models. While it was expected that these companies directly involved in the oil sector would be the best candidates to contain Sentiment in their reports that would correlate with oil prices, the study results proved otherwise.

Future research might be able to conclusively confirm this studies hypothesis. This future work may include:

- Utilizing multiple Sentiment analysis models. This study used Microsoft's cloud sentiment analysis tool at face value and was not benchmarked. Other Sentiment tools and models might produce better results.
- Using a different population of companies in the study. This might include both increasing the number of companies to include, or selecting companies from other sectors of the economy.
- Better understanding how the writing styles of different companies might influence Sentiment, and compensate for the same.
- Research different model that may be used to predict or explain commodity prices from Sentiment. In this study only linear regression models were considered.

References

Azmi Shabestari, M. & Romero, J. A. (2022). Textual Analysis and Future Performance: Evidence from Item 1A and Item 7. *Journal of Accounting, Auditing & Finance*.

Pablo, A (2009) - Sentiment Analysis in Financial News. [unpublished Academia.edu article]. *Harvard College*.

Xing, F., Malandri, L., Zhang, Y., & Cambria, E (2020). Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets. In *Proceedings of the 28th international conference on computational linguistics* (pp. 978-987).

Yang B, & Datta, A. (2014). Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science, 60*(6), 1371–1391.

Gupta, R, & and Chen, M. (2020) "Sentiment Analysis for Stock Price Prediction," *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* pp. 213-218, doi: 10.1109/MIPR49039.2020.00051.

Gao, L., Kampas. D., & Rinne, K, (2018) "Capturing Investor Sentiment: Advancing Predictability in Finance with Computer Science Approaches," *2018 IEEE 20th Conference on Business Informatics (CBI)*, pp. 97-99, doi: 10.1109/CBI.2018.10052.

Seals, E, & Price, S.R. (2020) "Preliminary Investigation in the use of Sentiment Analysis in Prediction of Stock Forecasting using Machine Learning," *2020 SoutheastCon*, pp. 1-2, doi: 10.1109/SoutheastCon44009.2020.9368258.

Rolnicki, S. G. H. (2018). *Applications of Natural Language Processing to Measure and Assess Risk in the U.S. Financial System*. (Publication No. 10841700) [Doctoral dissertation, Illinois Institute of Technology]. ProQuest Dissertations Publishing.

Feldman, Govindaraj, S., Livnat, J., & Segal, B. (2009). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915–953.

Soong, G. H. & Tan, C. C., (2021) "Sentiment Analysis on 10-K Financial Reports using Machine Learning Approaches," 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET), pp. 124-129, doi: 10.1109/ICSET53708.2021.9612552

U.S. Securities and Exchange Commission, "How to Read a 10-K," (2011). [Online]. Available: https://www.sec.gov/fastanswers/answersreada10khtm.html.

Wang, & Wang, X. (2012). Deceptive Financial Reporting Detection: A Hierarchical Clustering Approach Based on Linguistic Features. Procedia Engineering, 29, 3392–3396.

StockTwits. https://stocktwits.com/

Tebyanian and F. Hedayati, "Intelligent Crude Oil Price Forecaster," (2014) *2014 13th International Conference on Machine Learning and Applications*, pp. 453-455, doi: 10.1109/ICMLA.2014.79.

How to Read a 10-K. [How to Read a 10-K (sec.gov)]