JOURNAL OF DYNAMICS AND CONTROL

VOLUME 8 ISSUE 9

SENTIMENT AND MARKET ANALYSIS OF ENERGY SECTOR USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

Aryan Singh, Mainak Ghosh, Kumaran. M, A. Sharmila

School Of Electrical Engineering
Vellore Institute of Technology
VIT University, Vellore - 632014, Tamil Nadu, India

SENTIMENT AND MARKET ANALYSIS OF ENERGY SECTOR USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

Aryan Singh[1], Mainak Ghosh[2], Kumaran. M[3], A. Sharmila[4*]
School Of Electrical Engineering
Vellore Institute of Technology
VIT University, Vellore - 632014, Tamil Nadu, India
[1]aryansingh2019@vitstudent.ac.in,[2] mainakiplll@gmail.com ,[3] kumaran.m2023@vitstudent.ac.in, [4] asharmila@vit.ac.in

*Corresponding Author

ABSTRACT - *Crude Oil is an essential natural resource that has a big impact on the world economy, and sentiment analysis or opinion mining can often be tricky, especially when it comes to detection of emotions from news excerpts when they may be diplomatically phrased- we as humans tend to express ourselves differently and one sentence alone can have detections of different polarities. The focus of this paper is to propose an oil price trend prediction method by inferring sentiments from online news and social media. This is done by comparing the working and accuracies of the three most popular models used for opinion mining- VADER, Text Blob and Multinomial Naive-Bayes. Our aim is to analyze market sentiments using text recognition by comparing the accuracies of the two NLP-based models- VADER and Text Blob, and the Naive-Bayes model based on machine learning. While sentiment alone cannot always predict changes in commodity prices, with the help of technical analysis tools, better insights can be gained to determine in this case, the spot prices of Crude Oil.*

INDEX TERMS—Naive Bayes, VADER, Text blob, NLP, Crude Oil, Sentiment Analysis.

1. Introduction

Taking consideration of the commodity market in a world where the internet didn't play as big a role in every aspect of our lives as it does now, rates were determined according to supply and demand. The dominant factor was simple economics- supply greater than demand slashed prices and the opposite sent them through the roof. This project involves an investigation of the impact of consumer sentiments on the spot price of crude oil. The energy sector is a complex system and to estimate the effect that sentiment articulated in news and text extracted from social media applications have on the prices of crude oil, tools like the Natural Language Processing (NLP) based VADER and Text blob, also the Ma- chine Learning based Naive Bayes model have been used, with the results of each compared to the other and their efficiencies analyzed. To conduct this analysis and calculate the impact of consumer sentiment on the commodities market, a collection of data from the recent past was aggregated using web scrapping. VADER, Text blob and Naive Bayes were then implemented on the same datasets having stochastic reviews and newspaper articles, the sentiments of which are sorted into their polarities namely- positive, negative, or neutral for all the three algorithms. With news articles and tweets scraped from the web as our studying samples, our research aims to leverage the argument that the ups and downs in the energy sector- Crude Oil, in our case, are affected by relevant online information.

2. Related Research

A lot of research papers have been written in the field of sentiment analysis and its application in various industries. These research papers have studied sentiment classification, situation-based sentiment analysis, sentiment lexicons, sentiment analysis of social with tools like Tweepy, and specific industries such as finance, politics and healthcare. The development of sentiment lexicons like the WordNet Affect Lexicon and the SentiWordNet was a major focus of many of the early works on sentiment analysis. Other studies researched machine learning methods for sentiment analysis, including decision trees, support vector machines, and naive Bayes classifiers. Convolutional neural networks and recurrent neural networks are two examples of deep learning approaches to sentiment analysis that have gained popularity more lately. In a study done by Oussalah, M. [2], used sentiment analysis to see the impact of twitter on crude oil prices. The study found that sentiment on twitter regarding crude oil was a significant factor of price change and that sentiment analysis will become a major factor for market forecasting and volatility management in the crude oil industry. Sentiment analysis was employed in a different study by Kharde, V. [3] to assess Twitter users' opinions on crude oil businesses. According to the poll, people have generally negative attitudes towards the firms that produce crude oil. The most common concerns were worries about the environment, pricing, and geopolitical challenges. Tan, s. [5] proposed a feature selection technique for Naive Bayes in sentiment analysis and achieves higher accuracy than other traditional machine learning methods. Overall, these studies demonstrate the potential of sentiment analysis in the crude oil sector for market forecasting, risk management, customer sentiment analysis, social media monitoring, and sustainability analysis. In a market study done by Kilian, L. and Park, C. [9] the authors compared the relation between the crude oil sector and stock market, it was concluded that these returns are complex to bracket under one umbrella but majority of it depends on the state of global economy.

With the implementation of sentiment analysis industry practitioners can understand the market emotion.

3.Methods

To implement sentiment analysis of the crude oil sector using web scraping of data from Twitter and news articles, we used the following design structure approach as you can see in figure 1:

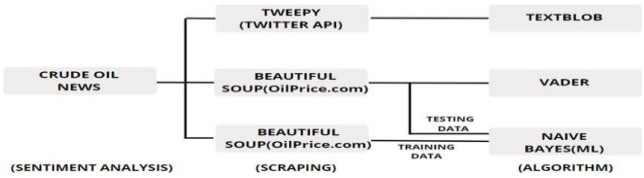
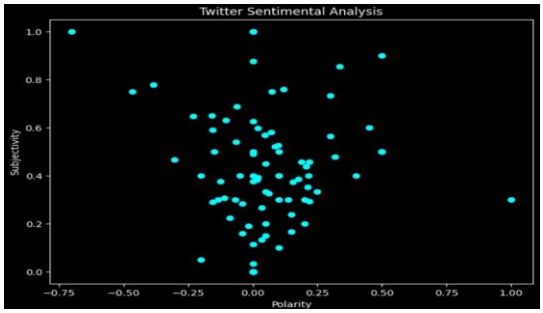


Figure 1: Entire design Flow Chart

3.1. Textblob

For implementing TextBlob in sentiment analysis of crude oil sector via Twitter API, the first step was to use the Twitter API to extract relevant tweets containing keywords or hashtags related to crude oil. After this we analyzed the tweets using pre-built sentiment analysis model of TextBlob, which gave a polarity score to each tweet showing whether it is positive, negative, or neutral. A Naive Bayes classifier and a lexicon-based approach are used in conjunction with pattern- based rules and machine learning approaches to determine the polarity score. In addition to the polarity score, TextBlob also provides subjectivity scores indicating the degree of subjective language used in the tweet. This was used for distinguishing between objective and subjective opinions in the analysis, as explained in figure 2. Finally, the polarity scores were aggregated and analyzed to gain insights into public opinion, sentiment trends, and other relevant metrics related to the crude oil sector. Although they are connected, subjectivity and polarity are not the same thing. A sentence may be subjective without having a certain polarity or it may be subjectively neutral but still have a strong positive or negative sentiment. The subjectivity score in the TextBlob algorithm determines how opinionated the statements are. Pertaining to the context of our project, the graph here shows the magnitude of polarizing nature of the crude oil market. In figure 2, x-axis shows the polarity and y-axis shows the subjectivity, most of the data points are in the range of negative and neutral polarity. As few of the European countries have started lifting sanctions on Russia people on social media have a neutral opinion of the market, The removal of sanctions will eventually result in lower oil commodity prices, which will lower inflation and, ultimately, improve conditions for the common citizen.

igure 2. Textblob Graph



3.2. VADER

For VADER the same approach of web scraping is done through beautiful soup and data is extracted from OilPrices.com. Most Machine Learning approaches involving NLP (Natural Language Processing) make use of a training set of data, whereas a lexical approach such as VADER relies on a combination of a dictionary which maps emotions based on a sentiment score, and the five heuristics- punctuation, capitalization, change in polarity due to the use of 'but', degree modifiers and examination of the tri-gram, which refers to a set of three lexical features. While both Text blob and VADER libraries produce fairly similar results, VADER picks up slang, emoticons and capitalization of words or letters. VADER calculates a score generally in the range of -4 to 4, but using Hutto's method for normalization:

$$\frac{X}{\sqrt{X2+\alpha}} \ldots\ldots (1)$$

- where *x* is the sum of sentiment scores of the input text.
- And *Alpha* is normalization parameter

We get a range of -1 to 1 for sentiment analysis of input text. Using basic math and stats functions in the python library Pandas

and VADER, the degree of polarity- positive, negative or neutral in each sentence of the input text is detected. With the help of Matplotlib library, a graph summarizing polarities of all news articles over time in the dataset is presented. A graph summarizing polarities of all news articles over time in the dataset is shown in figure 3. In this example we calculated the sentiment of around 50 articles related to crude oil, as we can see 20 articles gave a negative polarity and 16 articles gave a neutral polarity. This shows that the average sentiment in the crude oil sector is volatile as there is not much difference between the neutral and negative counts. Our analysis is based on the year 2023 specifically from the month of January to March and the model achieved a accuracy of 85 % which is higher than the industry range of VADER.

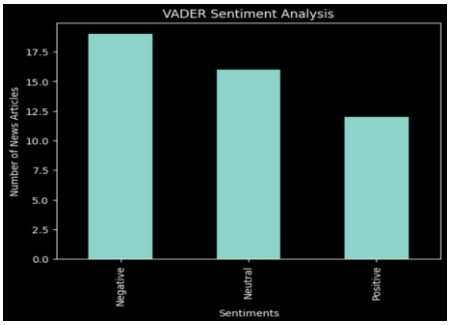


Figure 3. VADER Graph

3.3. Naive Bayes

The data is extracted via beautiful soup library from Oil- Price.com and converted into a csv file. The next objective was to label the news so that it can be trained properly. The better the training data set, the better results the naive bayes algorithm produces. By importing a dataset containing text data that needs to be categorized as positive, negative, or neutral sentiment, Excel is used to extract data for sentiment analysis. The dataset is divided into rows, with each row representing a text document or tweet, and columns holding pertinent information such as text content, date, and heading. By adding a column to the dataset and manually giving a sentiment label to each text document, the training data can be labelled. This labelling is important for the training of naive bayes algorithm, table 1. shows how a training data set would ideally look like.

Table 1. Data-Set Labelling

Sr. No.	News	Sentiment
1	The war between Russia and Ukraine could eliminate 1 million barrels of oil demand	Negative
2	Oil Is Likely to Stick To Its Production Schedule	Positive
3	Big Oil Is Investing Significant Resources in Renewable Energy Production.	Positive
4	Venezuela Lying About Its Oil Production	Negative
5	why is Latin America's oil industry collapsing	Negative
6	Analysts Are Unaware of Where the Real Trend in Oil Demand Is	Neutral
7	Iran Moves Millions of Barrels on Tankers As It Prepares for Sanctions Lifting	Positive

This training dataset was converted into a feature matrix for every sentence as shown in table 2. By converting a collection of text documents to a matrix of token counts, Every feature acts as a unique token. In this example, the first four news articles are taken from table 1. The feature sets are independent of each other and all of them have a corresponding dependent variable which tells us the final sentiment. [-1] represents negative sentiment and [1] represents positive sentiment. This feature set was used to train the dataset for implementing our naive Bayes model. However, implementation of TF-IDF is necessary before employment of the Naive Bayes model.

Table 2. Future Matrix

Oil	War	Lying	Production	Investing	Sentiment
1	1	-	-	-	-1
1	-	-	1	-	1
1	-	-	1	1	1
1	-	1	1	-	-1

TF-IDF stands for term frequency-inverse document frequency, and this is an algorithm which assigns weighting factors used to get the important features from the documents.

$$TF = \frac{\text{No.of times to appear in document}}{\text{No.of words in document}} \ldots\ldots (2)$$

$$IDF = \text{Log}\left(\frac{\text{Total No.of docs}}{\text{No.of docs with term t in it}}\right) \ldots\ldots (3)$$

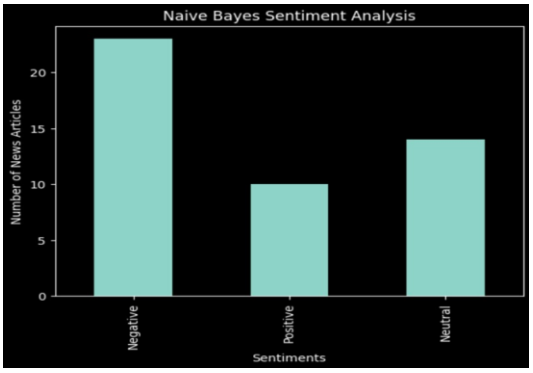
$$TF{:}IDF = TF * IDF \ldots\ldots (4)$$

Frequently occurring words do not actually influence the sentiment of the article, but they are predominantly about the subject that we are talking about. So, for example, if there are thousands of articles about smartphones, the word smartphone or its related tools and features are likely to occur frequently in every single article, which does not necessarily mean that feature or word will have any influence on the sentiment. If we do not penalize the high-frequency words and implement them in the machine learning model, it will ignore the other features and perform one on one mapping with the most frequently occurring word, which in our case happens to be- crude oil. This is where the mathematical construction TF-IDF comes up and penalizes the most frequently occurring words, as we can see in table 3. where we have taken the first news article from Table I into consideration for understanding of the algorithm.

Table 3. TF-IDF

Oil	War	Lying	Production	Investing	Sentiment
0.072	1	-	-	-	-1

After implementation of TF-IDF feature matrix, we trained the Naive Bayes classifier to perform sentiment analysis. Naive Bayes is a probabilistic model that makes predictions based on the conditional probabilities of each feature given in each class label. To train the classifier, the data was split into training and testing sets, then fit the model to the training data, and evaluate its performance on the testing data using the trained GaussianNB classifier. For testing, the same dataset was used as VADER to compare the results.



Both the algorithms gave a high negative polarity towards the crude oil sector, as we built our own dataset for naive bayes algorithm we achieved a F1 score of 0.9 with high accuracy. Tan, S. [5]'s paper concluded that Naive Bayes is an effective approach for sentiment analysis of natural language text when combined with feature selection techniques and our TF-IDF implementation helped the model train with more precision. If we compare figure 3 and figure 4. we can see that the energy sector is volatile in nature and as the VADER algorithm gave a similar sentiment, neutral aspect of the naive bayes algorithm helped us confirm it.

Figure 4. Naive Bayes Graph

4. Market Volatility Analysis

To reduce the vehement presence of volatility in crude oil market, identification of the magnitude or the extent of volatility is quintessential. There are multiple methods to measure the volatility of a sector. One common method is to use the standard deviation of the stock's log returns over a specific period; in our case we have used the data for the last 2 years. The daily standard deviation was transformed into an annual- aliased volatility. The annual volatility of the stock indicates how much its price is anticipated to move annually. The likely- hood of a stock's volatility increases with its level of volatility. Here in figure 5, x-axis shows us the log returns and y-axis show the frequency. Through the above-mentioned process volatility was calculated, which happens to be 42.05%. A considerable high volatility which shows the market is moving fast and may be affected by certain geopolitical factors. As per in our case Russia - Ukraine war is a major factor.

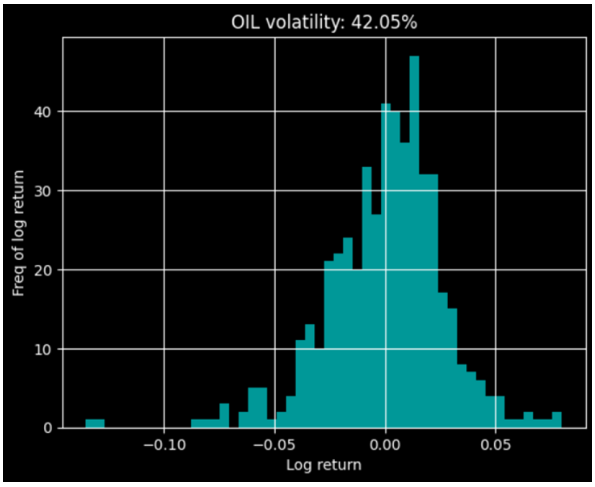


Figure 5. Market Volatility Graph

5. Results

The models are assessed using various indicators such as precision, recall, and f1 score as shown in table IV. Overall, the research paper can provide a comprehensive comparison of three popular sentiment analysis algorithms in the context of energy sector. This analysis provides a valuable insight into public and market opinion of the crude oil sector. The ability of these models to represent the intricate and subtle nature of human language has shown encouraging outcomes. The models evaluate the effectiveness of sentiment analysis in the energy sector, including the accuracy of sentiment analysis models, the utility of derived insights, and the impact of sentiment analysis on decision-making and global economy. Precision is a measure of how often a sentiment rating is correct.

$$\text{Current sentiment rate} = \frac{\text{Num.of Correct Queries}}{\text{Total Num.of Queries}} \quad \ldots\ldots (5)$$

We used this to check the overall accuracy of the system. For precision calculation we had to also calculate False Measure, the formula is:

$$\text{False Measure} = (+\text{ve}) \text{ false measure} + \text{Neutral FM} + (-\text{ve}) \text{ FM} \quad \ldots\ldots (6)$$

Subtract False Measure from the total number of queries extracted to get the correct number of queries, then we calculate precision using formula number 5.

Recall measures how many of the instances in the dataset were correctly predicted by the model. This could be seen as how accurately the system determines neutrality.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad \ldots\ldots (7)$$

The F-Score combines recall and precision. This is one of the most crucial metrics that will let you know how well your system is working. The formula for calculating F1 Score is:

$$\text{F1-score} = 2 * \frac{\text{Precision} * \text{recall}}{\text{Precision+recall}} \quad \ldots\ldots (8)$$

Table 4. Precision, Recall and F1 Score

Model	Precision	Recall	F1 score
Textblob	76%	52%	0.61
VADER	85%	78%	.81
Naive Bayes	91.6%	89%	.9

5. Conclusion

With our inbuilt dataset for naive bayes algorithm we got a high accuracy and F1 score of 91.6 % and .9 respectively, Vader had a decent accuracy of 85 % which helped us compare the market sentiment effectively and come to considerable conclusions on the high volatility of the market as discussed in the methodology section. Text blob had a F1 score of .61 and accuracy of 76 % which helped us analyze the subjective and objective difference between the tweets to see how factually correct the sentiment of the market.

If we compare the industry standards, prediction accuracy of Text Blob is around 68.8 % and for Vader it ranges between 70-80 %. Both the models performed better than the usual range of accuracy and us in built naive bayes algorithm outperformed by giving a F1 score of 0.9. Oil price estimates globally have become unclear because of conflicting supply and demand considerations. Fears of a recession have grown in both the United States and Europe, potentially reducing the demand for oil. The activities of the world's central banks, which have regularly raised interest rates to combat rising inflation, as well as China's economic slowdown, have also influenced the possibilities of dropping commodity demand. Furthermore, ongoing restrictions against Russian oil exports, most recently a 60 Dollars price cap on Russian crude oil imposed by the G7, have raised concerns about supplies from the world's second-largest supplier. Russia and China will be the primary countries driving the oil price projection in 2023, as the balance between tightening Russian supply and China's demand growth will drive oil prices.

Overall, sentiment analysis research has significantly advanced the field of natural language processing. By tracking public opinion on crude oil prices and projecting future price changes, we were able to employ sentiment analysis to uncover new patterns in the crude oil market.

Acknowledgement

With immense pleasure and deep sense of gratitude, we wish to express sincere thanks to our supervisor and guide Dr. Sharmila A, Professor, SELECT School, Vellore Institute of Technology, without her motivation and We are grateful to Vellore Institute of Technology, for motivating us to carry out research in the University and for providing us with infrastructural facilities and many other resources needed for the project. We wish to extend our profound sense of gratitude to our parents for all the sacrifices they made during the research and provide moral support and encouragement whenever required. Last but not the least we would like to acknowledge the support rendered by our colleagues in several ways throughout the project work.

References

[1] Hutto, C. and Gilbert, E., 2014, May. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

[2] Oussalah, M. and Zaidi, A., 2018, July. Forecasting weekly crude oil using Twitter sentiment of US foreign policy and oil companies' data. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 201-208). IEEE.

[3] Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

[4] Gujjar, J.P. and Kumar, H.P., 2021. Sentiment analysis: Textblob for decision making. Int. J. Sci. Res. Eng. Trends, 7(2), pp.1097-1099.

[5] Tan, S., Cheng, X., Wang, Y. and Xu, H., 2009. Adapting naive bayes to domain adaptation for sentiment analysis. In Advances in Information Retrieval: 31st European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31 (pp. 337-349). Springer Berlin Heidelberg.

[6] Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), pp.1093-1113.

[7] Hussein, D.M.E.D.M., 2018. A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), pp.330- 338.

[8] Wang, S.I. and Manning, C.D., 2012, July. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 90-94).

[9] Kilian, L. and Park, C., 2009. The impact of oil price shocks on the US stock market. International economic review, 50(4), pp.1267-1287.

[10] Filis, G., Degiannakis, S. and Floros, C., 2011. Dynamic correlation between stock market and oil prices: The case of oil-importing and oil-exporting countries. International review of financial analysis, 20(3), pp.152-164.