Text mining of news articles for prediction oil stock price

Tashbayeva A A

Kazakh National Technical University after K.I.Satpaeva, Almaty

Corresponding author's e-mail: a.tashbayeva@stud.satbayev.university

Abstract

At the present time, oil prices forecasting is one of the most issues, research areas. Most of the work involved in predicting oil prices was based on structured data, while in this work news articles would be used as unstructured data to sentiment analysis. The sentiment analysis is used to extract key information from texts and will be viewed by three perspectives of: negative, neutral, and positive sentiment. Finally, this work will analyze various views to improve prediction and get more accurate model.

Keywords: text mining, prediction stock price, sentiment analysis, NLP

1 Introduction

Oil for the country isn't only a product, but also the foundation of stability with which economic prospects are connected. Thus, fluctuations in oil prices have a crucial role for domestic economic stability. However, it's impossible to search out fundamental influencing factors on the change in oil prices, since geopolitics, market speculation may affect the change price of oil. Research has shown that price fluctuations are non-linear, and chaotic [1], which complicates the task of predicting oil prices.

With the event of the web and large data technologies, unstructured data began to be used more and more often, which store potential information, providing a replacement source of information for forecasting. To prove the considerable contribution of text mining to plug price forecasts, one can cite the instance of the work of Liu et al., who extracted a system of indicators from the company's Twitter to research its relationship with stock returns, and therefore the results show that Twitter indicators and stock prices are better connected than traditional industrial indicators [2].

For a deeper analysis of forecasting, it is not enough to simply extract the "quantity" in news, but the mood of the texts also plays an important role. A study conducted by Tetlock showed that media pessimism has predictive power for exchange prices [3]. Also, Li et al. used the LDA theme model and CNN neural network model to extract the mood of the news text, which improved the prediction model [4]. Thus, researches show that news emotions can better predict oil prices.

Based on the foregoing, we propose a new model for forecasting using text mining. We use the textual opinion obtained using textual analysis in a predictive model to spot the most effective thanks to use text. First, we examine the connection of text with news headlines and oil prices. Then, we examine the difference between the categories of textual moods to determine the effect on the worth. Finally, we show how an extra source of information can improve the forecasting result.

2 Overview

In this paper, the model is split into two main branches.

- The primary branch is that the processing of text data
- The second is that the processing of oil price data.

The ultimate model of oil price forecasting is combined with textual information on oil price and, finally, its effectiveness is evaluated.

3 Methods

This paper uses Brent crude oil price data (USD/barrel) from 27 March 2015 to 27 March 2020 as data. We select the data from 27 March 2015 to 27 March 2019 as training and modeling data and data from 28 March 2019 to 27 March 2020 as test data.

Based on the above price data, we draw a time series diagram which describes how oil prices fluctuate over time, are shown in Figure 1.

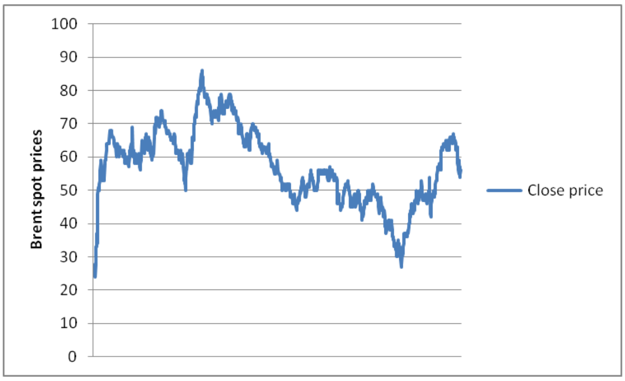
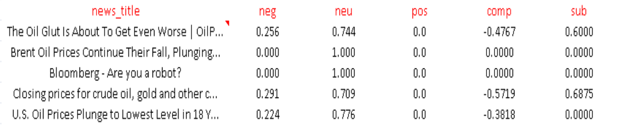


Figure 1 Daily Brent crude oil prices

In terms of web text, we use Python to scrape web text based on oil price related key words such as 'oil price', 'brent crude oil' from reliable online media such as Google News (https://www.news.google.com). We will pre-process the text data so as to urge more accurate results at the analysis stage. To do this, we filter the wrong text, and then we delete the abnormal words.

The next step is text analysis. For text analysis we use VADER. VADER is rule-based unsupervised method [5]. Due to this method, we get sentiment analysis of text defined as Figure 2



news_title	neg	neu	pos	comp	sub
The Oil Glut Is About To Get Even Worse | OilP...	0.256	0.744	0.0	-0.4767	0.6000
Brent Oil Prices Continue Their Fall, Plunging...	0.000	1.000	0.0	0.0000	0.0000
Bloomberg - Are you a robot?	0.000	1.000	0.0	0.0000	0.0000
Closing prices for crude oil, gold and other c...	0.291	0.709	0.0	-0.5719	0.6875
U.S. Oil Prices Plunge to Lowest Level in 18 Y...	0.224	0.776	0.0	-0.3818	0.0000

Figure 2 Sentiment analysis of news title

Our last step is to create a model. When forecasting, we chose logistic regression, RF (random forest) and decided to use a neural network in addition.

4 Conclusion

An additional source of data, namely textual information, provides significant advantages in forecasting oil prices. When forecasting, you need to consider that the model is better if the sentiment analysis of text is strong enough. Therefore, it is important to identify the strength of sentiment of text.

References

[1] Panas E, Ninni V *Are oil markets chaotic? A non-linear dynamic analysis*

[2] Liu L, Wu J, Li P, Li Q *A social-media-based approach to predicting stock comovement*

[3] Tetlock P C *Giving content to investor sentiment: The role of media in the stock market*

[4] Li J, Xu Z, Xu H, Tang L, Yu L *Forecasting Oil Price Trends with Sentiment of Online News Article*

[5] Hutto C J, Gilbert E *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Tex*