

Oil Prices Forecasting

Anton Kozlov, Sergey Skuredin, Daniil Shirshin

May 2024

Abstract

The topic and main question of our research is to determine the possibility of building a machine learning system that could predict market prices. We define our task as follows: to build a model that, based on information from news articles, will be able to predict the behavior of market prices for oil.

Our team, through discussion, chose a problem whose solution could be of interest to each team member.

This work is aimed at consolidating and demonstrating the skills acquired after mastering theory and performing practical exercises in the field of Natural Language Processing, as well as exploring the possibility of practical application of such skills.

1 Introduction

The choice of task is primarily related to the personal interest of each team member. Each of us agreed that the solution experience and the results obtained can be used in our professional activities.

However, the issue under consideration lies on the surface and could be interesting both from the point of view of determining the very possibility of predicting market prices based on news information, and applying the resulting model in the practical sphere, for example, when generating news or determining the potential for financial investment.

Our work is an attempt to build a model that can be in demand both by the authors themselves and by other researchers when solving a similar problem.

The uniqueness of our work is related to one of the key issues of machine learning - the selection and preparation of initial data for training.

1.1 Team

Our team has three members who did the following work:

1. **Anton Kozlov:** created task, parsing news, prepared dataset, discussion.
2. **Sergey Skuredin:** created task, prepared report, discussion.
3. **Daniil Shirshin:** prepared pipeline, learning model, discussion.

2 Related Work

Other researchers have been engaged in similar projects related to predicting prices for various market goods (securities, precious metals, commodities) based on open-source news information.

The project “STONKS” from a Russian student of the NLP training program from MIPT in 2023 was considered [1].

In this work, the author chose a model trained with reinforcement, and also prepared an agent that learned to perform actions in the environment using information from the model. The environment consisted of graphs of stock prices in a given time domain of observation and current market values. The agent’s actions had three possible meanings - buy, sell, or remain out of the market.

The goal is to maximize the reward, which is achieved by not exceeding a given divergence distance predicted by the model and the new state of the environment.

As input, the model takes environmental data and returns a probability distribution for specific actions and subsequent repetitions.

The author tried to use FCN (fully connected neural network) and LSTM (long short-term memory) as the basis of his model.

To evaluate the model, the author used several metrics, including his own development. An interesting result was the conclusion about the advisability of using reinforcement learning for complex problems, including what is confirmed by one of the metrics adopted to evaluate the model - S3PS (Stock Possible Profit Percent Score).

Other studies also explored hybrid approaches combining machine learning with sentiment analysis for crude oil price forecasting. For instance, Hu et al. (2018) proposed a hybrid machine learning approach that leverages investor sentiment data for forecasting crude oil prices [2].

Additionally, Jiang et al. (2021) developed a forecasting model based on Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks for crude oil price prediction [3].

3 Model Description

The model is based on the Russian-language BERT - “sbert”, which seems to be the largest transformer for the Russian language (in terms of the number of parameters). Thus, we already have pre-trained embeddings of words on a huge corpus of text, and our task mainly comes down to training the classifier.

The model is characterized by the presence of classical BERT (as shown above) and a fully connected linear layer for the classification task.

For training purposes, we provide news articles and associated pricing information as input. We get class labels for changes in price behavior. This is how representations are formed for the classifier, which in turn selects one of three states - the price will rise, the price will fall, the price will not change.

4 Dataset

To solve the problem of predicting oil prices, we needed to collect a data set with news articles and, in fact, oil prices. The data must be related to each other by a time parameter.

To prepare the data set, the relevant news site “neftegaz.ru” was selected, which contained about 10 thousand articles. Based on the parsing results, the first part of the data set with news and timestamps was formed. An example is shown in Fig. 1.

	link	datetime	headline	description	articleBody	date
0	https://neftegaz.ru/news/finance/831520-np2-da...	2024-04-27 13:06:01+00:00	РетроChina почти месяц не могла выгрузить парт...	А. Данготе хотел больше прибыли, покупая нефть...	Китайская государственная энергетическая компа...	2024-04-27
1	https://neftegaz.ru/news/finance/831527-tseny...	2024-04-27 09:01:23+00:00	Цены на нефть выросли по итогам предыдущей сес...	За минувшую неделю цена Brent увеличилась на 2...	Цены на нефть завершили неделю в плюсе: WTI/25...	2024-04-27
2	https://neftegaz.ru/news/transport-and-storage...	2024-04-26 13:31:27+00:00	Индия вновь принимает нефть из России на танке...	Владимир Тихонов, принадлежащий Совкомфлоту, н...	Нефтеналивной танкер, который принадлежит нахо...	2024-04-26
3	https://neftegaz.ru/news/transport-and-storage...	2024-04-26 11:35:11+00:00	Чехия в 2025 г. будет получать нефть из 20 гос...	С 2025 г. по трубопроводу TAL в Чехию, будет п...	Чехия в 2025 г. после расширения мощностей тра...	2024-04-26

Figure 1: Sample of first part of the dataset - “news”

For price data, the site “nasdaq.com” was selected, from which a csv file was downloaded, also with a timestamp. This is how the second part of the data set is formed. An example is shown in Fig. 2.

	Date	Close/Last	Open	High	Low	date
0	2024-04-30	87.86	88.32	88.71	87.47	2024-04-30
1	2024-04-29	88.40	89.25	89.30	88.10	2024-04-29
2	2024-04-26	89.50	89.21	89.83	88.80	2024-04-26
3	2024-04-25	89.01	88.11	89.26	87.29	2024-04-25
4	2024-04-24	88.02	88.46	88.80	87.65	2024-04-24

Figure 2: Sample of second part of the dataset - “prices”

In the next step, the data was combined over time. At the same time, for news released on non-trading days, when there is no trading information, we assigned the values that corresponded to the previous trading days. An example of the combined data is presented in Fig. 3.

The preparation of the data set did not stop at combining pieces of information. To solve the problem, we decided to preprocess and label the data set. For this we used the following aids:

- NLTK library and its “stopwords” module;
- The pymorphy3 library and its module “MorphAnalyzer”;
- Helper library tqdm.

	date	headline	description		article	close	open	high	low
0	2024-04-27	PetroChina почти месяц не могла выгрузить парт...	А. Данготе хотел больше прибыли, покупая нефть...	Китайская государственная энергетическая компа...		89.50	89.21	89.83	88.80
1	2024-04-27	Цены на нефть выросли по итогам предыдущей сес...	За минувшую неделю цена Brent увеличилась на 2...	Цены на нефть завершили неделю в плюсе.\n\n26 ...		89.50	89.21	89.83	88.80
2	2024-04-26	Индия вновь принимает нефть из России на танке...	Владимир Тихонов, принадлежащий Совкомфлоту, н...	Нефтеналивной танкер, который принадлежит нахо...		89.50	89.21	89.83	88.80
3	2024-04-26	Чехия в 2025 г. будет получать нефть из 20 гос...	С 2025 г. по трубопроводу ТАЛ в Чехию, будет п...	Чехия в 2025 г. после расширения мощностей тра...		89.50	89.21	89.83	88.80
4	2024-04-26	Нефть дорожает и завершает неделю в плюсе	За неделю Brent подорожала на 2,4%, WTI - на 0...	Цены на нефть растут:\n\n25 апреля 2024 г. сто...		89.50	89.21	89.83	88.80

Figure 3: Sample of the dataset - “news and prices”

Russian stop words were taken from the NLTK library. Using the pymorphy3 library, we lemmatized the text into words and counted their total number. Next, based on the considered project “STONKS”, the idea was used to calculate the most popular negative and positive words. Having carried out an analysis using the expert method of the first thousand (1000) most frequently occurring words, we identified two classes of positive and negative words, meaning a corresponding increase and decrease in oil prices. The result is presented in Table 1.

Thus, we divided the data set into two classes - positive news, meaning an increase in the price of oil, and negative news, meaning a fall in the price of oil. Table 2 shows the parameters of the generated data set.

5 Experiments

5.1 Metrics

Due to the fact that the main task of the project is related to classifier training, we use the CrossEntropyLoss loss function.

5.2 Experiment Setup

Our model training experiment included the following steps to post-process the input data:

1. The input data that makes up the article is combined into one text.
2. We calculate the price for the day in which the news article appeared, that is, we determine its change (Eps == \$0.05).
3. Based on the calculation, a label indicating an increase, decrease, or no change in prices is assigned.
4. The prepared data set is randomly divided into parts for training and testing in a ratio of 70 to 30.

Type	Positive words	Negative words
Count	46	55
Values	рост, увеличить, вырасти, начать, начало, хороший, прибыль, восстановление, перспектива, превысить, запустить, очередной, первый, принять, рекордный, стратегический, успешно, возможность, результат, экономика, высокий, максимум, позитивный, оптимизм, эффективность, стабильность, стабильный, поддержка, прирост, темп, призвать, эффект, уникальный, значительно, финансирование, уверенно, лидер, серьёзный, уверенный, комплексный, разрабатывать, инвестиционный, усиление, продукция, развивать, ключевой	аварийная ситуация, аварийный, авария, взрыв, вниз, война, дефицит, дефицитный, дешеветь, жертвы, забастовка, загрязнение, запрет, инцидент, катастрофа, конфликт, кризис, минус, нарушение, нарушитель, негативный, неоднозначный, неожиданный, неопределённость, несчастный случай, нехватка, обвал, ограничение, опасение, падать, падение, пожар, проблема, протест, разлив, резкий, рецессия, риск, санкции, слабо, снижение, сокращение, спад, трагедия, трудный, убыток, угроза, утечка, ущерб, хмао, штраф, экологическая катастрофа, экологический ущерб, экономический кризис

Table 1: Positive and negative words

News	10028
Positive news	7637
Negative news	2391
Vocabulary size	9629
Date range	2016/08/30 – 2024/04/27

Table 2: Statistics of the dataset

Next, the model was trained with the main training parameters of the main model and Baseline, which are presented in Table 3.

Type	Our main model	Baseline
Embedding size	256	256
Batch size	4	8
Epochs	5	3

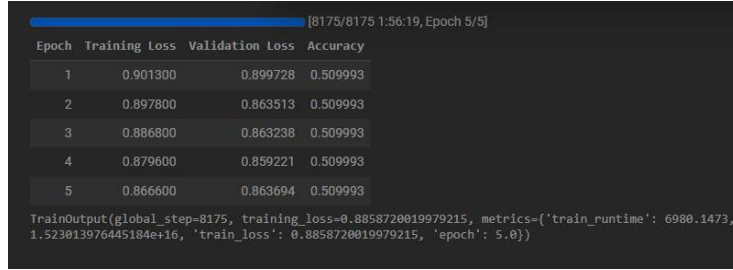
Table 3: Characteristics of models

5.3 Baselines

For Baseline, we used BERT embeddings reduced by a factor of 10 (in terms of the number of parameters) and a fully connected linear layer.

6 Results

Using our main model, the following results were obtained (Fig. 4):

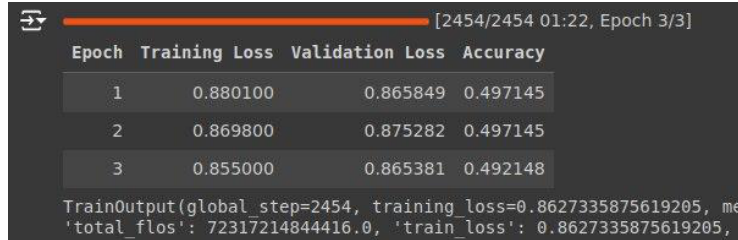


Epoch	Training Loss	Validation Loss	Accuracy
1	0.901300	0.899728	0.509993
2	0.897800	0.863513	0.509993
3	0.886800	0.863238	0.509993
4	0.879600	0.859221	0.509993
5	0.866600	0.863694	0.509993

TrainOutput(global_step=8175, training_loss=0.8858720819979215, metrics={'train_runtime': 6980.1473, 1.523813976445184e+16, 'train_loss': 0.8858720819979215, 'epoch': 5.0})

Figure 4: Result of our main model

The following results were obtained on the Baseline model (Fig. 5):



Epoch	Training Loss	Validation Loss	Accuracy
1	0.880100	0.865849	0.497145
2	0.869800	0.875282	0.497145
3	0.855000	0.865381	0.492148

TrainOutput(global_step=2454, training_loss=0.8627335875619205, metrics={'train_runtime': 82.0, 'total_flos': 72317214844416.0, 'train_loss': 0.8627335875619205, 'epoch': 3.0})

Figure 5: Result of Baseline model

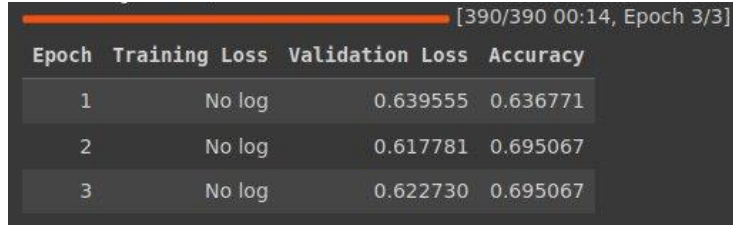
Results of the main model and those associated with the use of the normalized dataset are shown in Fig. 6:

7 Conclusion

First of all, it should be noted that even the Baseline model showed results close to random guessing. The main model learned a little better, which can be justified by a larger transformer (10 times more parameters) and a better tokenizer, meaning that the model depends on the quality of the representation.

However, with different options for parameters, we noticed that the size of embeddings has virtually no effect on the final result.

The results show a weak relationship between news articles and subsequent changes in oil prices.



[390/390 00:14, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	No log	0.639555	0.636771
2	No log	0.617781	0.695067
3	No log	0.622730	0.695067

Figure 6: Result of main model with normalized input dataset (last extension of experiments)

We believe that embeddings trained on a data set that is more relevant to a given task than those trained on non-specific information could improve predictive accuracy. Alternatively, a simpler way to improve the model could be to expand existing representations by pre-training for our specific task.

References

- [1] Sirazhetdinov, R. (2023). STONKS - Stock Market Forecasting Using News Articles With Reinforcement And Sequence Learning.
- [2] Hu, Z., Zhao, Y., Hua, W., Lu, Z., Kang, S. (2018). *A Hybrid Machine Learning Approach for Forecasting Crude Oil Prices Using Investor Sentiment Data*. Journal of Computational Science, 28, 362-370.
- [3] Jiang, S., Liang, Y., Ding, Z. (2021). *Crude Oil Price Forecasting Model Based on Long Short-Term Memory Network and Gated Recurrent Unit Network*. Applied Soft Computing, 108, 107483.