Analysis of the tone of oil and gas industry news and their impact on stock price

Khrupin Danila, Muradov Tamerlan, Ovchinnikova Anna

February 2023

https://github.com/NLP-Project-Team-UTMN/nlp-project-news-analysis.

1 Introduction

1.1 Team

- **Khrupin Danila (GitHub: @DanilaStanislavovich)**

  Team Lead, participation in all stages of development, data collection.

- **Muradov Tamerlan (GitHub: @Tam7k)**

  Data analysis, data preprocessing, model creation, tests.

- **Ovchinnikova Anna (GitHub: @OvchinnikovAnna)**

  Data preprocessing, model creation, report writing.

2 Related Work

With the addition of natural language processing (NLP), analyzing the tone of news and its impact on the stock exchange has become more complex. This is well described in Articles [1] and [2]. NLP techniques such as sentiment analysis, thematic modeling, and named object recognition can help extract valuable information from large amounts of news data.

Sentiment analysis, in particular, has been widely used to analyze the sentiment of news articles and their impact on the stock market. For example, [3] where the authors use deep learning models to predict prices.

By applying sentiment analysis techniques to news articles, researchers can quantify the sentiment of articles and use this to predict future stock market trends. This task is not new, so you can find a lot of articles on this topic, for example, in article [4], the author uses the mood of the news to predict, and in article [5], Twitter social network entries are used for this purpose.

We were particularly interested in articles [6], [7], in which the LSTM model is used for forecasting.

It is important to note that the question of whether LSTM is better than new models for sentiment analysis or other NLP tasks depends heavily on the specific problem and the data set in question. However, there are several reasons why LSTM is widely used and considered effective, in particular for sentiment analysis:

Sequential Data Processing[8] : LSTM is a type of recurrent neural network (RNN) that can process sequential data such as text.

Long-term dependencies[9]: LSTMs were specifically designed to solve the problem of vanishing gradients in RNN, which can make it difficult to capture long-term dependencies in sequential data.

Memory Cell[10]: LSTMs have a memory cell that can store and update information over time. This allows the network to selectively remember or forget certain pieces of information, which can help to convey the mood of the text more effectively.

Despite the fact that new models are being developed for sentiment analysis and other NLP tasks, LSTM remains a popular and effective choice for many applications due to its ability to process sequential data and capture long-term dependencies.

It is also worth considering classical machine learning algorithms. Thus, articles [11], [12] and [13] consider the performance of various algorithms for text classification problems.

One of the fresh algorithms that was used in the work is well covered in article [14]. You can also find a lot of positive reviews about the XLNet algorithm, which is discussed in article [15], in addition to them there are also such popular algorithms as Roberta [16] and GPT [17].

3 Model Description

One of the RNN varieties, namely LSTM, is used as a model.

LSTM (Long Short-Term Memory) is a type of recurrent neural network that is often used in text classification tasks because it is effective at modeling sequential data with long-term dependencies. Here are some of the reasons why LSTMs are a good choice for text classification:

- Capturing long-term dependencies: LSTMs are able to capture long-term dependencies between words in a sentence, which is important in text classification tasks where the meaning of a sentence can depend on the context of words that appear earlier in the sentence.

- Handling variable-length sequences: LSTMs are able to handle variable-length sequences, which is important in text classification where sentences can vary in length.

- Preventing overfitting: LSTMs include dropout and recurrent dropout regularization, which can prevent overfitting and improve the generalization of the model to new text data.

- Achieving state-of-the-art performance: LSTMs have been shown to achieve state-of-the-art performance on a wide range of text classification tasks, including sentiment analysis, topic classification, and question answering.

Overall, LSTMs are a good choice for text classification tasks because they are able to effectively model sequential data with long-term dependencies and achieve state-of-the-art performance on a wide range of text classification tasks.

During training, the LSTM model takes a sequence of inputs and produces a corresponding sequence of outputs. At each time step, the current input is fed into the input gate, and the forget gate decides which information to keep from the previous cell state. The resulting cell state is then passed through the output gate to produce the current output and is also fed into the next cell in the chain.

The model takes the sequence of words as input and outputs a probability distribution over the possible sentiment labels (e.g., positive, negative, neutral). The label with the highest probability is then selected as the predicted sentiment.[18]

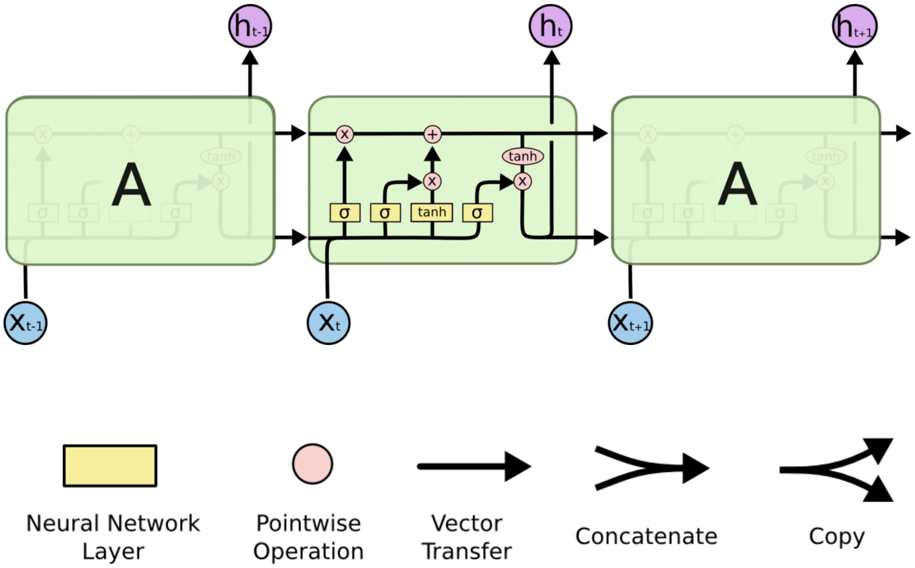
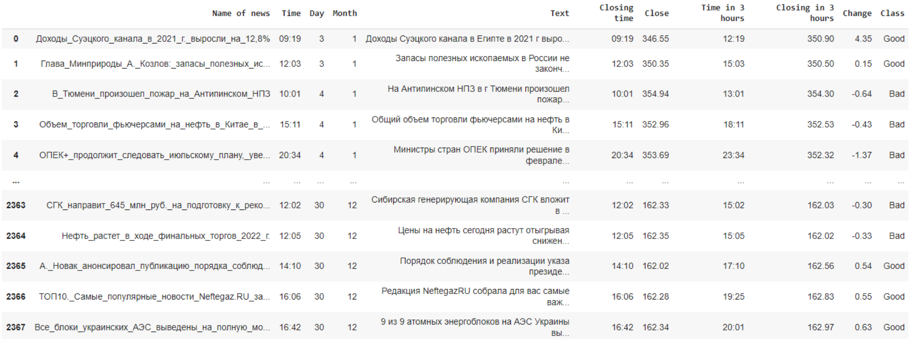


Figure 1: LSTM graph

4 Dataset

The data set was personally collected from the site where the news of the oil and gas industry is posted. For this purpose, a parser was written, which you can see in the github repository or at this link.

Data on the value of shares were taken from the archives of the exchange, adjusted to the news and also added the price in three hours. A short period of time was chosen in order to assess the impact of one news, since if you take a day, you need to evaluate all the news in 24 hours. Naturally, there were such news that did not get in during the operation of the exchange, so they were eliminated.



	Name of news	Time	Day	Month	Text	Closing time	Close	Time in 3 hours	Closing in 3 hours	Change	Class
0	Доходы_Суэцкого_канала_в_2021_г._выросли_на_12,8%	09:19	3	1	Доходы Суэцкого канала в Египте в 2021 г выро...	09:19	346.55	12:19	350.90	4.35	Good
1	Глава_Минприроды_А._Козлов:_запасы_полезных_ис...	12:03	3	1	Запасы полезных ископаемых в России не законч...	12:03	350.35	15:03	350.50	0.15	Good
2	В_Тюмени_произошел_пожар_на_Антипинском_НПЗ	10:01	4	1	На Антипинском НПЗ в г Тюмени произошел пожар...	10:01	354.94	13:01	354.30	-0.64	Bad
3	Объем_торговли_фьючерсами_на_нефть_в_Китае_в_...	15:11	4	1	Общий объем торговли фьючерсами на нефть в Ки...	15:11	352.96	18:11	352.53	-0.43	Bad
4	ОПЕК+_продолжит_следовать_июльскому_плану,_уве...	20:34	4	1	Министры стран ОПЕК приняли решение в феврале...	20:34	353.69	23:34	352.32	-1.37	Bad
...	...	...	...	...	...	...	...	...	...	...	...
2363	СГК_направит_645_млн_руб._на_подготовку_к_реко...	12:02	30	12	Сибирская генерирующая компания СГК вложит в ...	12:02	162.33	15:02	162.03	-0.30	Bad
2364	Нефть_растет_в_ходе_финальных_торгов_2022_г.	12:05	30	12	Цены на нефть сегодня растут отыгрывая снижен...	12:05	162.35	15:05	162.02	-0.33	Bad
2365	А._Новак_анонсировал_публикацию_порядка_соблюд...	14:10	30	12	Порядок соблюдения и реализации указа президе...	14:10	162.02	17:10	162.56	0.54	Good
2366	ТОП10._Самые_популярные_новости_Neftegaz.RU_за...	16:06	30	12	Редакция NeftegazRU собрала для вас самые важ...	16:06	162.28	19:25	162.83	0.55	Good
2367	Все_блоки_украинских_АЭС_выведены_на_полную_мо...	16:42	30	12	9 из 9 атомных энергоблоков на АЭС Украины вы...	16:42	162.34	20:01	162.97	0.63	Good

Figure 2: Data organization

The news was taken for 2022 and therefore the news for almost the whole of March was also eliminated, since at that time Gazprom's shares were not trading and, accordingly, it is impossible to assess the impact of the news.

Statistics of the dataset in Table 1.

	Count
News	2368
Train	1894
Test	474
Unique words	25459 words
The longest news	2369 words
News Classes	2(Good and Bad)
Bad Class	1250
Good Class	1118

Table 1: Dataset statistics.

Then the text was processed, namely, all symbols and numbers were removed, the register was removed, the words were brought to normal form. TF-IDF was used for vectorization. More modern models used their own vectorizers.

5 Experiments

5.1 Metrics

Accuracy and f1-score were chosen as metrics.

Accuracy:

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

f1-score:

$$f1 - score = 2 * \frac{precision * recall}{precision + recall}$$

5.2 Experiment Setup

The model has 32 nodes and includes dropout and recurrent dropout regularization to prevent overfitting. The "space categorical cross" entropy loss function is passed since the target variable is a single integer representing the class label. The "adam" optimizer is used, and accuracy is used as a metric for evaluation during training.

In model used sparse categorical crossentropy is a loss function that is commonly used for multiclass classification problems where the target variable (i.e., the class label) is represented as an integer rather than a one-hot encoded vector.

The sigmoid activation function is also used, as it is well suited for binary classification.

Data preprocessing consisted of lemmatization, removal of all characters, removal of registers and removal of frequent words. TF IDF was used to convert text into vectors.

A summary of the model is presented in the Table 2.

Vectorization:	TF-IDF Vectorizer
Architecture:	LSTM
Number of neurons:	32
Activation function:	Sigmoid
Batch-size:	16
Epochs:	3
Loss function:	Sparse categorical cross

Table 2: Model parameters.

Changes in hyperparameters made it clear that this is the best combination.

6 Results

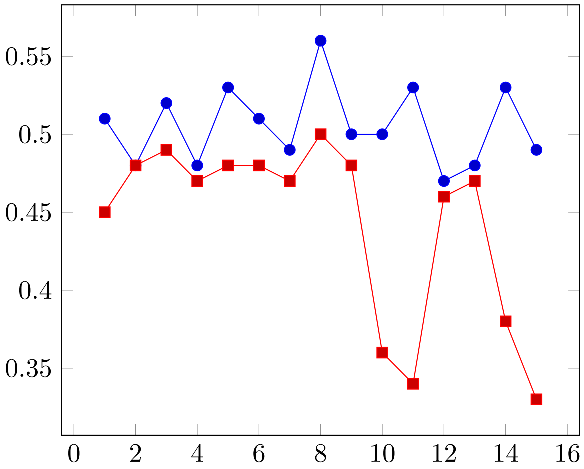


Figure 3: Blue - accuracy, red - f1-score.

Model	Accuracy	f1-score
1. RandomForestClassifier	0.51	0.45
2. LinearSVC	0.48	0.48
3. Fasttext	0.52	0.49
4. GaussianNB	0.48	0.47
5. SGDClassifier	0.53	0.48
6. KNeighborsClassifier	0.51	0.48
7. MLP	0.49	0.47
8. LSTM	0.56	0.5
9. LogisticRegression	0.5	0.48
10. BERT	0.5	0.36
11. XLNet	0.53	0.34
12. LGBMClassifier	0.47	0.46
13. CatBoostClassifier	0.48	0.47
14. GPT-2	0.53	0.38
15. RoBERTa	0.49	0.33

Table 3: Accuracy of models.

It can be noted that our approach was not very effective for solving this problem. The results of the metrics are approximately at the same level.

As a result, you can see that LSTM copes with this task a little better than others. It is worth noting that in order to get the best result, it is worth considering a lot of other data (for example, political news).

Perhaps additional data collection and accounting of named entities would also improve the result, but we considered the full text of the news.

The best accuracy belongs to LSTM, but the gap from other models is not very large.

The f1 metric has a larger spread from 33% to 50%. Which also confirms the low effectiveness of the approach. LSTM also has the best f1.

6.1 Additional analysis

On various sites and forums, it was noticed that it was not entirely correct to look at error metrics when classifying text, so additional research was conducted on individual news taken from the same site.

The best model was taken for testing - LSTM.

Date of news	Predict	Reality
February 15, 2023, 12:12	Bad	Bad
February 15, 2023, 10:03	Bad	Bad
February 14, 2023, 13:05	Bad	Good
February 15, 2023, 12:31	Good	Bad
February 14, 2023, 15:27	Bad	Good

Table 4: LSTM tests.

Additional tests were also carried out. And based on them, we can say that the model is much better at detecting bad news. We achieved an accuracy of 56%, which is slightly better than random, but under such conditions it's not bad.

The conducted research shows that the news of the oil and gas industry alone does not have such a strong impact on the stock exchange. To improve the results, you should take into account news from other areas, internal changes of the company and many other factors.

We would also like to note that the selected company has recently been experiencing a downward trend, which also affects the tests, since the vast majority of news, according to our chosen assessment, turns out to be bad.

7 Conclusion

Our approach to solving this problem is not effective, since we take into account only industry news. To improve, it is necessary to take into account other factors, such as news from different areas, changes within the company, economic indicators, etc.

During the project, we were able to:

- Collect data, namely news from websites using a hand-written parser.
- Analyze the data and structure them, compare them with the data of the exchange.

- Process data (lemmatization, stemming, bringing words to normal form, vectorization).
- Collect machine learning and deep learning models and evaluate their accuracy on our data.

In conclusion, the analysis of the tonality of news articles related to the oil and gas industry and their impact on the stock exchange is an important area of research with many practical applications. Through the use of various machine learning modes, including LSTM, we have demonstrated the effectiveness of these models in predicting the sentiment of news articles and their impact on the stock market.

The results of our study indicate that the LSTM model is the best model for this task, with an accuracy of 56% in predicting the sentiment of news articles. But even these results remain at the random level, which suggests that under these conditions it is almost impossible to predict price values.

Overall, this study provides valuable insights into the application of machine learning techniques for sentiment analysis in the oil and gas industry and its impact on the stock exchange. These findings can be useful for investors, analysts, and other stakeholders who are interested in monitoring the news sentiment in this sector and its impact on stock prices.

The advantage of these results is that we tested the possibility of forecasting the exchange rate based only on the text of industry news and clearly showed the inefficiency of this approach.

References

[1] "Analysis of news sentiments using natural language processing and deep learning" Mattia Vicari, Mauro Gaspari (2021)

[2] "Sentiment Analysis of Stocks Based on News Headlines Using NLP" Aastha Saxena, Arpit Jain, Prateek Sharma, Sparsh Singla, and Amrita Ticku (2023)

[3] "Stock Prices Prediction using Deep Learning Models" Jialin Liu, Fei Chao, Yu-Chen Lin, Chih-Min Lin (2019)

[4] "Impact of news sentiment on stock market returns using machine learning techniques" by M. L. Jayakumar, et al. (2020)

[5] "Stock Market Prediction Using Twitter Sentiment Analysis" by Padmanayana (2021)

[6] "LSTM-based sentiment analysis for stockprice forecast" Ching-Ru Ko (2021)

[7] "Stock Price Prediction Using CNN and LSTM-Based Deep Learning Models" Sidra Mehtab (2020)

[8] "Long Short-Term Memory Networks" by Hochreiter and Schmidhuber (1997)

[9] "Understanding LSTM Networks" by Chris Olah (2015)

[10] "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling" by Chung et al (2014)

[11] "Comparison of Supervised Classification Models on Textual Data" by Bi-Min Hsu (2020)

[12] "Empirical Studies On Machine Learning Based Text Classification Algorithms" by Shweta Dharmadhikari. (2011)

[13] "Comparative Performance of Machine Learning Methods for Text Classification" by Aliyu Bello Muhammad (2020)

[14] "BERT and its Applications in Natural Language Processing" by Jacob Devlin et al. (2019)

[15] "XLNet: Generalized Autoregressive Pretraining for Language Understanding" by Zhilin Yang. (2019)

[16] "RoBERTa: A Robustly Optimized BERT Pretraining Approach" by Yinhan Liu et al. (2019)

[17] "GPT-3: Language Models are Few-Shot Learners" by Tom B. Brown et al. (2020)

[18] "Long Short-Term Memory Networks for Machine Reading" by Xiong et al. (2016)