IDENTIFICATION OF NEWS TEXT CORPORA INFLUENCING THE VOLATILITY OF FINANCIAL INSTRUMENTS

A.S. Stankus [a]

Saint Petersburg State University, 7-9 Universitetskaya emb., Saint Petersburg, 199034, Russia

E-mail: [a] alexey@stankus.ru

Using neural networks to predict changes in financial markets is a promising task. For more accurate forecasting, it is necessary to determine the tone of the texts of the articles, whether the news carries positive or negative information for the market. Standard approaches to using pretrained neural networks aimed at analyzing user reviews are not successful due to the fact that professional reporters try to present their articles in a neutral way, which leads to incorrect conclusions. In this article, we will talk about the possibilities of training neural networks to analyze the sentiments of articles based on volatility data in the volatility of financial markets.

Keywords: Neural networks, attention-based, transformer, BI-LSTM, sentiments of articles, financial market

Alexey Stankus

1. Statement of the problem

Let's take the news collected from Reuters for a certain period and the oil price for the same period. Using one of the most advanced architectures of the BERT transformer [1-2], we get the following definition of tonalities [tab. 1]:

Table 1. The result of determining the sentiment of news

Data class	Number of articles	Percentages
Neutral	97152	61.55%
Negative	60206	38.14%
Positive	496	0.31%

From the obtained distribution of results, it is obvious that the model is carries most articles to the "neutral" class. This fact can be expected - news articles rarely contain a emotional component. In most cases, the authors adhere to a strict style, which is designed to present the dry facts. Standard models of news sentiment recognition are usually trained on short and emotional messages such as social networks posts or customers reviews. Nevertheless, it is considered how the labels relate to the price movement [tab.2]:

Table 2. The relationship between sentiment and price movement

News label	Price increase	Price decrease
Negative	30604	29602
Positive	272	224

As a result of calculating p-value = 0.07, we can conclude that there is no statistical significance of the obtained news breakdown. Thus, the use of the transformed model is not justified - it is trained to find the wrong relationships that are necessary to solve the problem posed within the framework of this work. Overfitting under the given conditions may also not lead to a positive result due to strong differences in both the training set and the predicted feature. The use of the above-mentioned models requires a complete learning process from scratch, which can be realized only in the presence of an extremely voluminous and correctly labeled training data array.

From this, it is necessary to make conclusions about the creating your own data markup with subsequent training of the neural network.

2. Selecting news and texts pre-processing

By the reaction of the course, it is possible to determine the presence of relevant information that carries a positive or negative value for asset owners. To do so we must proceed the following steps:

- Relying on volatility as an indicator of market expectations [3];
- Select the news preceding the event;
- Pre-processing texts;
- Train the neural network;
- Checking the result on predictions.

The first thing to consider is the average volatility that is characteristic of the market and associated with the opening and closing of exchanges around the world. [fig. 1].

Figure 1. Average volatility values by hour

To identify the moments of the market reaction to the information received, the deviation of the volatility in the time interval from the value obtained at the last step can be used. After obtaining the values of the volatility deviations, you can build an approximation of the first and second derivatives (v′, v″). Further, as the moments of anomalous market reaction, sharp jumps in the V″, accompanied by long-term preservation of the positivity of the V′, are considered. [fig. 2].

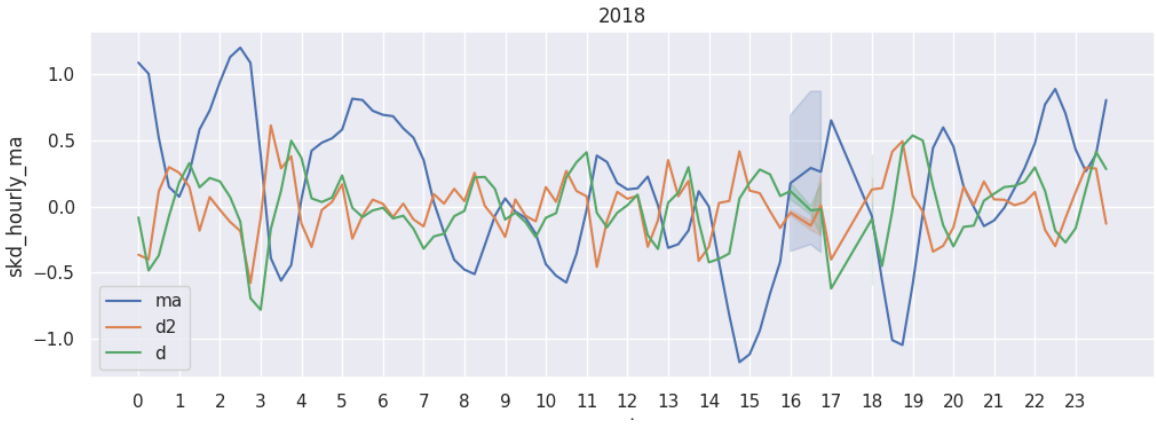

Figure 2. Moving volatility, its second and first derivatives

Next, we have to do text pre-processing. As part of the ongoing work, a large number of regular expressions and NLP packages have been applied to improve the quality of the input data. In particular, the following operations were performed on the texts of articles:

- replacement of html-mnemonics and special characters;
- censoring swear words;
- correction of obvious typos;
- removal of redundant punctuation marks;
- defining the language of publication and separating different languages from each other;
- addition of description texts with the text of the article, if necessary.

To speed up and improve the efficiency of training, the corpus of texts is filtered by keywords presumably relevant to the asset under study. We got 53 000 text news for each class.

3. Sentiment analysis model

In order to prevent overfitting [4], dropout layers have been added to the used neural network architecture, which randomly change the values of the previous layer (dropout) or disable some variables of the embedding layer (spatial-dropout). Neural network has following structure [tab. 3]:

Table 3. Model layers

Layer	Size	Parameters
Embedding	256	2560000
Dropout	256	0
B-LSTM	128	164352
B-LSTM	128	98816
Dropout	128	0
Dense	128	16512
Dropout	128	0
Dense	32	4128
Dense	1	65

After 45 epochs of GPU training on supercomputer “Govorun”, the result is [tab. 4]:

Table 4. Training results

Sample	categorical accuracy
Training	0.8657
Test	0.5374

Due to the use of a more complex three-class markup function, it should be noted that in this case the problem was solved not of a binary, but of a multiclass classification. When setting the problem of multiclass classification, the “basic” accuracy of the random number generator, which is the boundary of the meaningfulness of the result, is 1/3, and not 1/2, as in the case of binary classification, respectively.

4. Results

After training, we get the following results for texts assessment:

- «Russia committed holding round talks week the Belarussian capital Minsk ending violence eastern Ukraine, senior Kremlin aide said Monday»

Class	Confidence level
Neutral	0.005
Negative	0.003
Positive	0.992

- « Two people died at least 13 injured an explosion a factory belonging to Gulf Oil Corporation Ltd the southern Indian city Hyderabad, police said Monday »

Class	Confidence level
Neutral	0.003
Negative	0.904
Positive	0.093

- «Former world number Maria Sharapova cruised past Kazakhstan's Zarina Diyas into semi-finals the Shenzhen Open China Thursday»

Class	Confidence level
Neutral	0.96
Negative	0.03
Positive	0.01

In all these examples, the model classification results correspond to the real expected market reaction - the increase in oil exports by Saudi Arabia is assessed as negative news for the oil price, progress in resolving world tension in Ukraine is assessed as positive, and irrelevant news is assessed neutrally. However, such a review of the results cannot serve as a basis for drawing conclusions about the quality of the model. It is necessary to determine the effectiveness by additional verification.

References

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://arxiv.org/abs/1810.04805

[2] A. Vaswani, N.Shazeer, Niki Parmar et al. Attention Is All You Need. 2017

[3] A. Atkins, M. Niranjan, E.Gerding. Financial news predicts stock market volatility better than close price // The Journal of Finance and Data Science. 2018. Vol. 4, pp. 120-137.

[4] R. Pascanu, T. Mikolov, Y. Bengio. On the difficulty of training Recurrent Neural Networks. 2013.