Team members: Vilceanu Ovidiu

Mocanu Alexandru

Introduction:

The aim of this project is to establish whether there is a correlation between stock increase and decrease in terms of price, and the most popular headlines, according to Reddit, which were obtained by user voting.

Throughout history, many have tried to predict the stock market using various means, from pure statistical methods to complex pattern finding. We will try to find out how much the prices are influenced by popular news, for each day respectively, and point out which words carried the most weight when it comes to the prediction.

For implementation, we were thinking of two approaches : either use Naïve Bayes or try to tune Bert to do the classification. For the former we expect better results because it takes into account meaning of sequences of words, whereas Naïve Bayes only considers words individually.

At https://neptune.ai/blog/predicting-stock-prices-using-machine-learning we can see a introduction into Loss Functions, and a Moving Average model is presented to us (a model considering only the price, widely known in trading). It evaluates several such models, then it tries out a LSTM approach. Finally, it talks about models currently in development or alternative ways of correlating the data.

At http://www.ijmlc.org/vol7/614-A101.pdf we are presented with various popular models for such classification, such as Support Vector Machine, Naïve Bayes, K-nearest neighbors, ADA Boost, Neural Network and so on, and we are given a comparison of accuracies, based on the category of the input (bank, pharmacy stock etc.)

Repository : https://github.com/roger440/CVDL

Implementation

Components:

1. Setup

   This is the component we used in order to prepare the data for the BERT model to work with. The data is filtered, excluding the records with invalid format, and grouped into folders used for training and testing, each categorised in folders for classification

2. Main

   This is the part in which we load the data using a tensorflow data loader, download the 'Small Bert' model, configure the loss function, the optimiser, hper-parameters, and use the training set in order to

   fine-tune the already pre-trained BERT model. The validation set will be used in order to compute the loss for

   the current epoch. For each epoch, details regarding the loss value and the accuracy are printed

BERT:
Bert is a machine learning model used for solving natural language processing problems. It achieves this in 2 steps:
-pre-training: a phase in which BERT uses unsupervised learning in order to understand language and context
-fine-tuning: slightly modifying the parameters obtained in pre-training, uses supervised learning to corelate input to output

Findings, conclusions

The accuracy we obtained is rather inconclusive, in most configurations the model predicted the right output in ~51 % of cases. We could make the argument that we proved there is a correlation between news and stock prices, but the correlation seems to be rather small.

Possible reasons:

- more data was needed, we only trained the model with data from the past 3-4 years, not enough to fully understand the fluctuations in the market

- predicting the stock market requires data from different domains, such as politics, laws, natural events, and those cannot be extracted with accuracy from our dataset only

- an accuracy close to 100 % would suggest that the market is completely dependent on reddit news, and completely independent of any other exterior event. This hypothesis is rather absurd.