MARKETPSYCH RESEARCH

Deep Learning Prediction of the EUROSTOXX 50 with News Sentiment

Author of paper:
S. A. BOROVKOVA and M. M. T. DIJKSTRA, Vrije Universiteit Amsterdam, VU Amsterdam

Date of paper: 30 June 2018

Link: https://poseidon01.ssrn.com/delivery.php?
ID=658002065071097119106107070093067098030040014018086061111065113075071104018071091030030053038123014014001023066113075

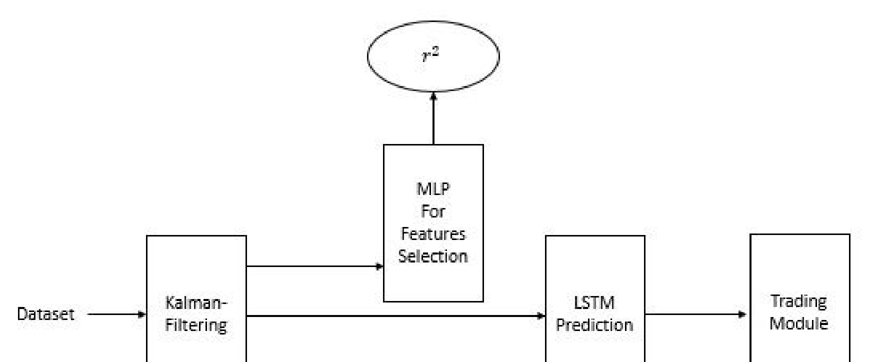
1 Proposed Methodology

1.1 Main Idea

A workflow is proposed to:

1. smooth the dataset using Kalman Filtering
2. select features for prediction using Multi-Layer Perception(MLP)
3. with selected features, predict the price for the next timestep using LSTM
4. simulate trading strategy based on LSTM predictions and compare it to strategies based on Price Momentum.

In the diagrams below, a box means a model while a oval means a evaluation metric that will be explained in later sections



Data Granularity
Data was aggregated at 5 minutes interval from 09:00 to 17:30 CET and there are 102 datapoints per day. Each timestep refer to a 5 minutes interval.

Sentiment Dataset
The sentiment data consist of:

1.Thomson Reuters News Analytics
2.TRMI

Table 1. Descriptive statistics of the aggregated news variables.

Feature name	Observations	Mean	Median	Std. Dev.	Skew.	Kurt.	Max.	Min.
$Relevance$	8772	0.60	0.59	0.36	-0.17	19.36	1	0.01
$Sent_{words}$	8772	223.71	133	252.45	2.99	11.54	2756	7
Tot_{words}	8772	674.23	541	615.22	2.75	15.20	6172	8
$Sent_{pos}$	8772	0.33	0.34	0.24	0.42	-0.89	0.86	0.02
$Sent_{neut}$	8772	0.18	0.20	0.18	1.36	1.36	0.93	0.11
$Sent_{neg}$	8772	0.28	0.26	0.27	0.44	-1.25	0.82	0.01
$Lnkd_1$	8772	10.13	6	12.29	2.90	12.94	125	0
$Lnkd_2$	8772	15.96	10	18.59	2.88	11.86	173	0
$Lnkd_3$	8772	34.75	21	42.23	3.24	15.10	367	0
$Lnkd_4$	8772	52.49	33	60.69	3.06	13.06	527	0
$Lnkd_5$	8772	75.11	49	82.63	2.86	10.91	593	0
$Item_1$	8772	2.01	1	4.02	4.22	21.09	36	0
$Item_2$	8772	2.83	1	5.75	3.60	13.66	38	0
$Item_3$	8772	5.56	1	14.59	4.31	19.68	106	0
$Item_4$	8772	8.04	1	21.37	3.82	15.00	165	0
$Item_5$	8772	11.47	2	31.55	3.65	12.70	178	0
$Buzz$	8772	13.09	10	11.59	3.19	19.37	173.7	0.2

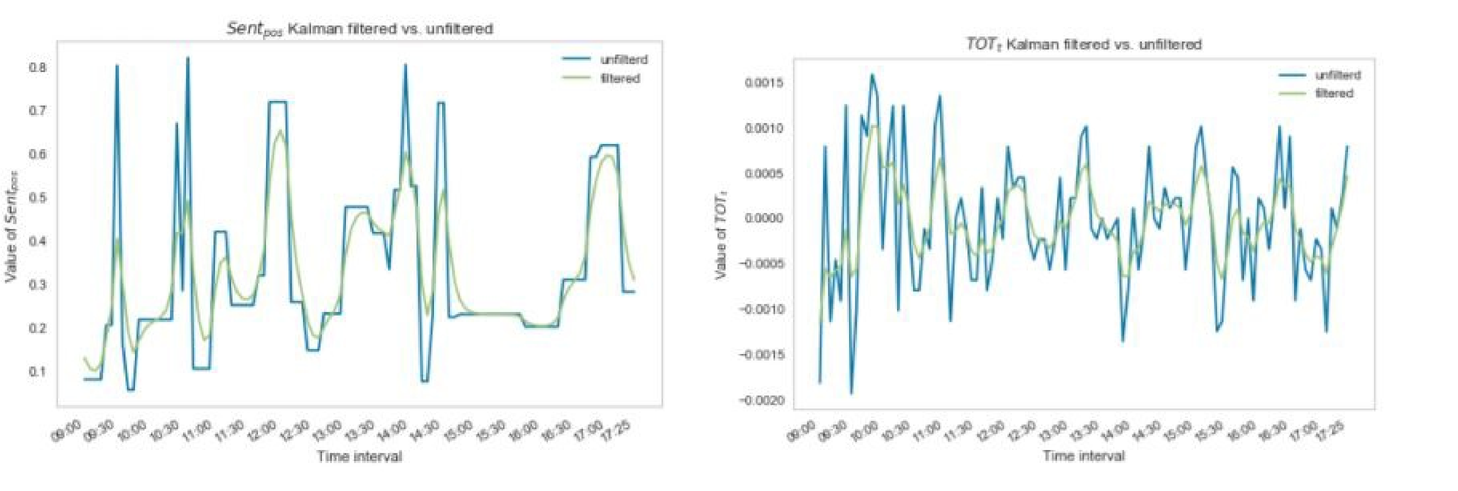
Financial Dataset
The financial dataset consist of prices of the EUROSTOXX 50 (SX5Et) and the prices of three large stocks in the index: Total (TOTt), Siemens (SIEt) and Sanofi (SANt).

Train-Validate-Test Split
Due to details not given, the train-validate-test split is approximately as follows:

3 months Train	10 days Validate	24/11/2016 - 30/11/201 Test

1.2 Smoothing of dataset using Kalman Filtering

The sentiment dataset was Kalman filtered before any input to any model to reduce noise, as shown below:

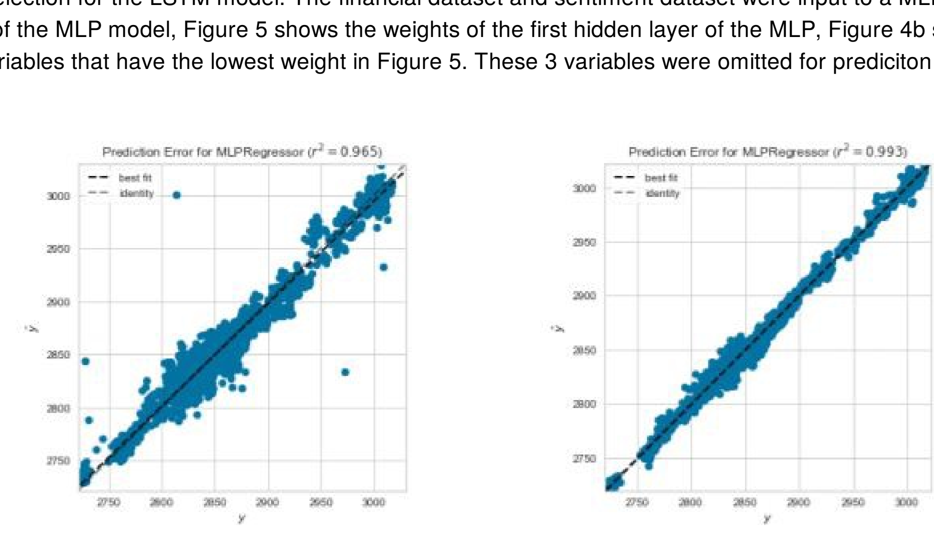


(a) (b)

Figure 1. Kalman filtered vs. unfiltered features $Sent_{Pos}$ (a) and Tot_t (b)

1.3 Select Features for prediction

A MLP was used for feature selection for the LSTM model. The financial dataset and sentiment dataset were input to a MLP to predict the target price (t+1). Figure 4a shows the fit of the MLP model, Figure 5 shows the weights of the first hidden layer of the MLP, Figure 4b shows the fit of the MLP model after removing the 3 variables that have the lowest weight in Figure 5. These 3 variables were omitted for prediciton of prices using LSTM in later steps too.



(a) (b)

Figure 4. Prediction error of the MLP Regressor for original input variables (a) and for adjusted input variables (b).

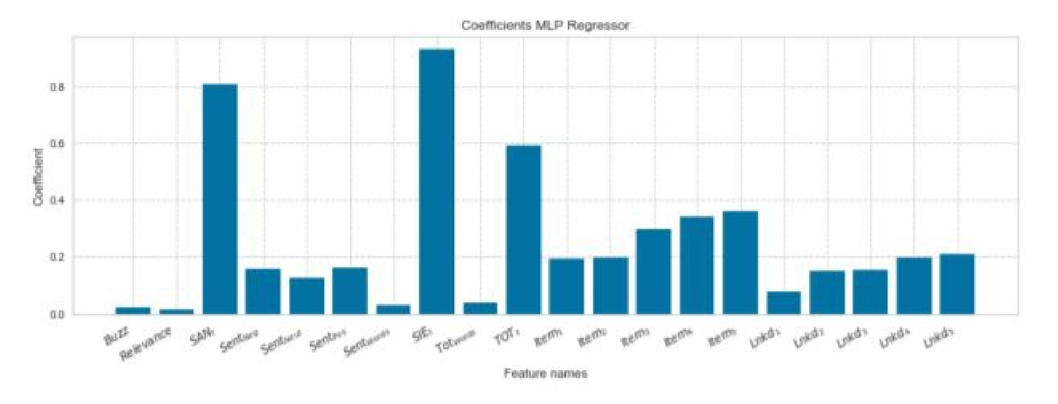


Figure 5. Absolute weight coefficients MLP for original input variables.

1.4 LSTM to predict price

After the features were selected in the previous step, the selected features were fed into a LSTM model. The LSTM model was configured to predict T+1 price given data at T.

1.5 Simulate and Compare Trading Strategies

After the LSTM predictions have been made, 4 trading strategies based on LSTM predictions and 4 strategies based on price momentum were simulated to compare the results. The 8 strategies are as follows:

Table 4. TRNA and TMRI feature details. *The history periods are by default given by 12 hours, 24 hours, 3 days, 5 days and 7 days prior to the time stamp of the news item.*

Name	Description
LSTM-I	Take long and short positions every 5 minutes based on the signal
LSTM-II	Sell positions are taken if the signal is -1, the position is closed out if the signal is 1
LSTM-III	Hold position if consecutive signals are the same
LSTM-IV	Only take positions if $0.0019 < \hat{\Delta p_y} < 0.0039$
Mom_{12}	Take long and short positions based on 1-hour momentum
Mom_{36}	Take long and short positions based on 3-hour momentum
Mom_{72}	Take long and short positions based on 6-hour momentum
Mom_{102}	Take long and short positions based on 1-day momentum

2. Results and Evaluation

2.1 Results of LSTM predictions

Mean Direction Accuracy
Mean Direction Accuracy (MDA) is used to indicate the percentage of forecasts that are predicted in the correct direction.
Mean Average Percentage Error
Mean Average Percentage Error (MAPE) is obtained to measure the error as a percentage of the target price.

Figure 10 and Table 2 show the results. The average MDA is 0.6255 and the average MAPE is 0.0942.

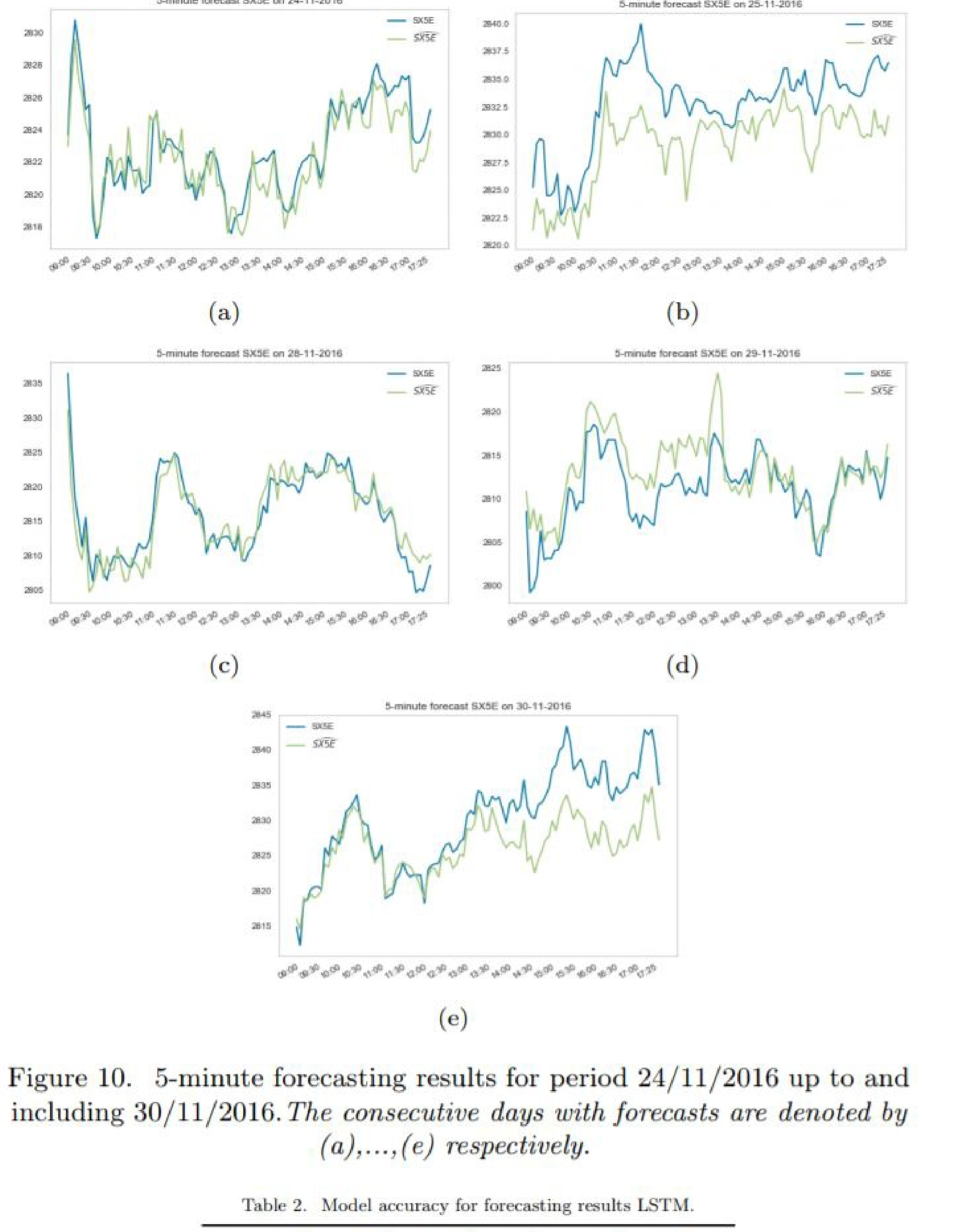


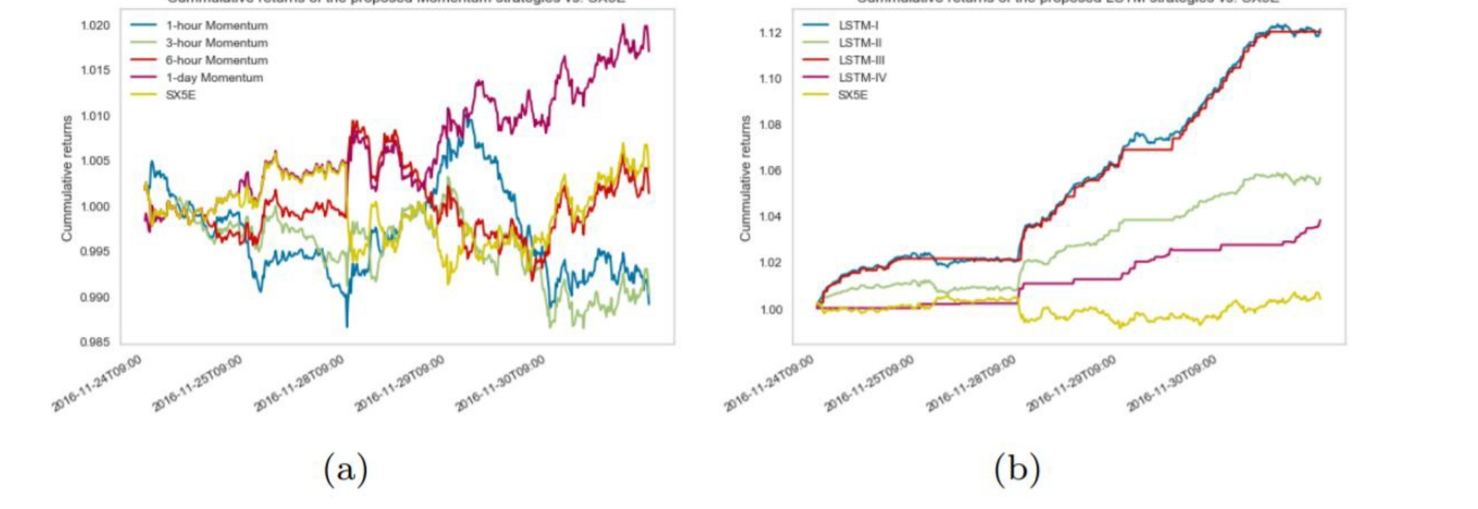
Figure 10. 5-minute forecasting results for period 24/11/2016 up to and including 30/11/2016. *The consecutive days with forecasts are denoted by (a),...,(e) respectively.*

Table 2. Model accuracy for forecasting results LSTM.

Date	24-Nov	25-Nov	28-Nov	29-Nov	30-Nov	Average
MDA	0.6747	0.5000	0.7156	0.6471	0.5980	0.6255
MAPE	0.0396	0.1274	0.0642	0.0934	0.1464	0.0942

2.2 Results of the Trading Strategies

The following shows the results of the 4 trading strategies based on LSTM predictions and 4 trading strategies based on Price Momentum.



(a) (b)

Figure 12. Cumulative returns of the proposed Momentum (a) and LSTM (b) strategies vs. $SX5E$.

3 Personal Takeaway

3.1 MLP for feature selection

A MLP before LSTM could be used to improve feature selection.

3.2 Prediction of Prices instead of Binary Classification

Prediction of numeric returns or prices instead of binary classification could be used to optimise the LSTM model.