Differential Privacy LSTM for Stock Prediction with Financial News

Zhaoyang Liu, Cecilia Yu, Emma Jin, Zonghan Li

Abstract

Inspired by the paper "*DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News*", this project aims to perform stock price prediction using various machine learning techniques and data processing steps. In addition to replicating the paper result from scratch, we processed news data into **sentiment features** while adding **Gaussian perturbation for differential privacy**, which enhances the robustness of the model and leads to more stable results by adding robustness against **adversarial examples**; we introduced a few more model classes (across **statistical methods, machine learning and deep neural nets**); finally, we applied **different time windows**, tested data using **multiple metrics**, and validated results on **multiple industry indexes** in terms of the asset price and returns. We observe and conclude that **the DP-LSTM model would provide better prediction accuracy and all models perform better on the return data instead of the price data**. The time window analysis shows us that **the lag of 15 days would lead to best prediction results as it captures enough variance and it is not over-fitted.**

Table of Contents

01 Data Processing

02 Model Selection

03 Observations

04 Conclusion & Improvements

01

Data Processing

Data Processing

Stock Index

We incorporated four stock market indices from Yahoo Finance API: **S&P 500, Nasdaq 100 Index, the Dow Jones Industrial Average, and the Russell 2000 Index**. Furthermore, we included the simple daily return of these stock indices as prediction variables.

Financial News Article

Four major news sources: CNBC, Fortune, Reuters, and Wall Street Journal

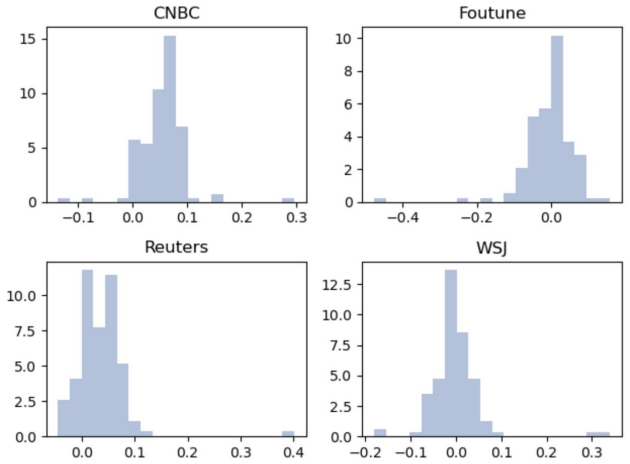
Different companies may have varying views and source data on similar market topics at a given time. In our model, each article sentiment score was considered a separate feature, as it explains different variance attributes to target prices or returns.

NLP (NLTK Vader)

1. How to Streamline the 6,714 Photos Cluttering Your Phone
2. Fire and Fury: Kimmel, Colbert Skewer Trump Over Wolff Book Fortune
3. This entrepreneur is ringing up sales restoring vintage telephones - MarketWatch

news1	how	to	streamline	the	6,714	photos	cluttering	your	phone		
news2	fire	and	fury	kimmel	colbert	skewer	trump	over	wolff	book	fortune
news3	this	entrepreneur	is	ringing	up	sales	restoring	vintage	telephones	marketwatch	

	compound	negative	neutral	positive
news1	0.0	0.0	1.0	0.0
news2	0.7269	0.404	0.596	0.0
news3	0.296	0.0	0.804	0.196



02

Model Selection

Model Selection

LSTM

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that was specifically designed to model non-linear, multi- featural data and long-range dependencies in sequences.

- No sentiment
- With sentiment
- Differential privacy

ARIMA

ARIMA (AutoRegressive Integrated Moving Average) is a statistical model that is particularly useful model for predicting linear, univariate time series data that display a pattern or trend over time

Random Forest

Random Forest models are particularly useful for handling non-linear and high-dimensional data and are not prone to overfitting.

- GridSearchCV method for tuning (lag)

03

Observations

Evaluation Metrics

Mean Absolute Error	Mean Squared Error	Accuracy	Mean Error Percent
$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$1-\frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$	$\frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$
measure of the average magnitude of difference between predicted and actual values	measure of the average squared magnitude of difference between predicted and actual values	measure the accuracy of a model's predictions as a percentage of the actual values	Measure of the average percentage difference between predicted and actual values

Model Performances - MSE

	No Sentiment	With Sentiment	DP	ARIMA	Random Forest
SP	335.6792	350.8661	313.0404	3004.0790	389.2055
SP_r	0.0000	0.0000	0.0000	0.0000	0.0001
NDX	1204.9574	1910.9455	1838.4121	28466.6088	44847.2560
NDX_r	0.0000	0.0000	0.0000	0.0001	0.0001
DJI	46692.1061	43289.7420	43728.1917	256693.9434	79804.7380
DJI_r	0.0001	0.0001	0.0001	0.0001	0.0001
RUT	285.1282	467.1597	374.4224	3093.7415	5246.3267
RUT_r	0.0001	0.0001	0.0001	0.0000	0.0000

Model Performances - MAE

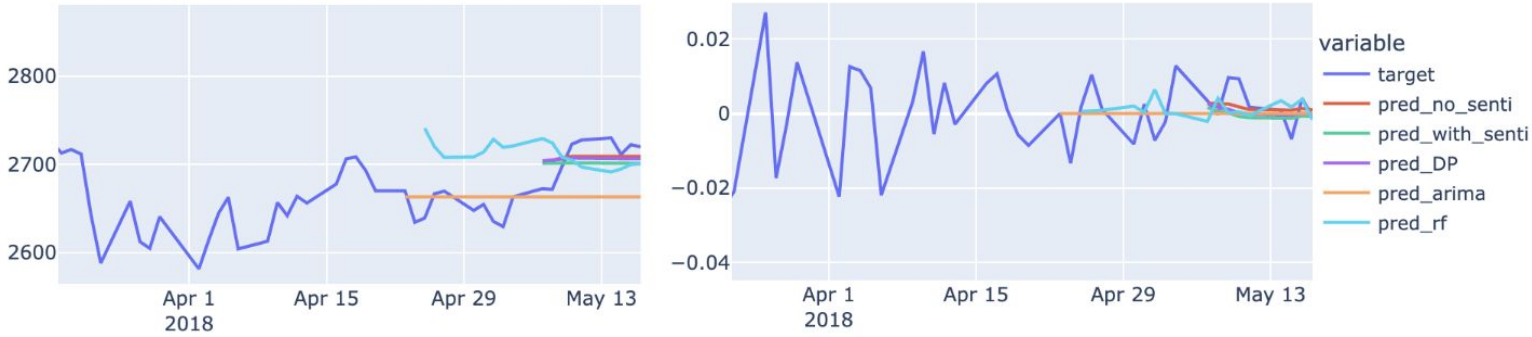
	No Sentiment	With Sentiment	DP	ARIMA	Random Forest
SP	16.6860	16.5129	16.2685	51.5201	14.7470
SP_r	0.0060	0.0058	0.0058	0.0053	0.0063
NDX	30.9512	37.1904	37.0674	158.0062	185.5751
NDX_r	0.0043	0.0043	0.0043	0.0057	0.0070
DJI	179.8912	164.2563	163.4007	469.6973	222.0697
DJI_r	0.0073	0.0068	0.0069	0.0061	0.0071
RUT	15.1944	18.8928	20.3630	52.2700	71.0966
RUT_r	0.0062	0.0059	0.0058	0.0051	0.0055

Model Performances - Accuracy

	No Sentiment	With Sentiment	DP	ARIMA	Random Forest
SP	0.9936	0.9938	0.9939	0.9811	0.9946
NDX	0.9956	0.9944	0.9946	0.9773	0.9733
DJI	0.9927	0.9935	0.9934	0.9810	0.9910
RUT	0.9907	0.9887	0.9890	0.9678	0.9562

Prediction v.s. Actual

S&P 500 Index & Return

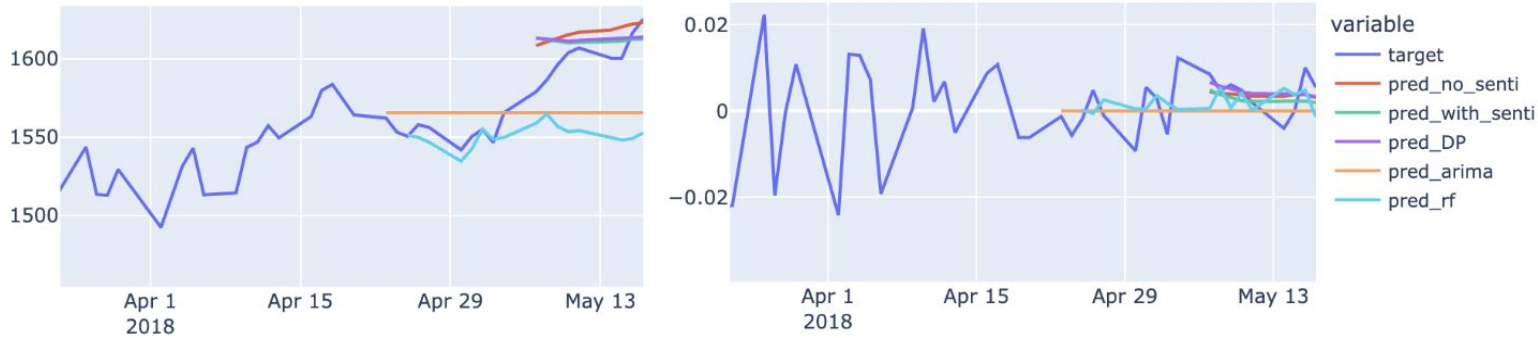


Nasdaq Index & Return

Prediction v.s. Actual

Dow Jones Industrial Average Index & Return

Russell 2000 Index & Return



Time Window Analysis

To explore the influence of the length of time windows on the prediction models, we set the time window to 3 days, 10 days, and 15 days, and examine the three LSTM models' performances on the S&P 500 Stock Index. Following shows the mean absolute error of the three models with different time windows.

	No Sentiment	With Sentiment	DP
3 days	16.7959	16.8809	16.3856
10 days	16.7605	16.8566	15.6518
15 days	16.6992	16.1918	15.3357

04

Conclusion & Improvements

Conclusion

The LSTM model outperformed ARIMA and Random forest for index time series prediction. All time series models work better on the return data as it is normalized with the same scale. LSTM with DP would achieve best performance in most scenarios as the features it takes would be more robust.

Future Improvements

The further improvement can be made through extending the sentiment datasets while utilizing Neural Networks would lead to better representation that has the huge potential for improvements. In addition, we could spend more time on model parameter tuning for ARIMA and Random forest that may enhance the performance.

Thank you!