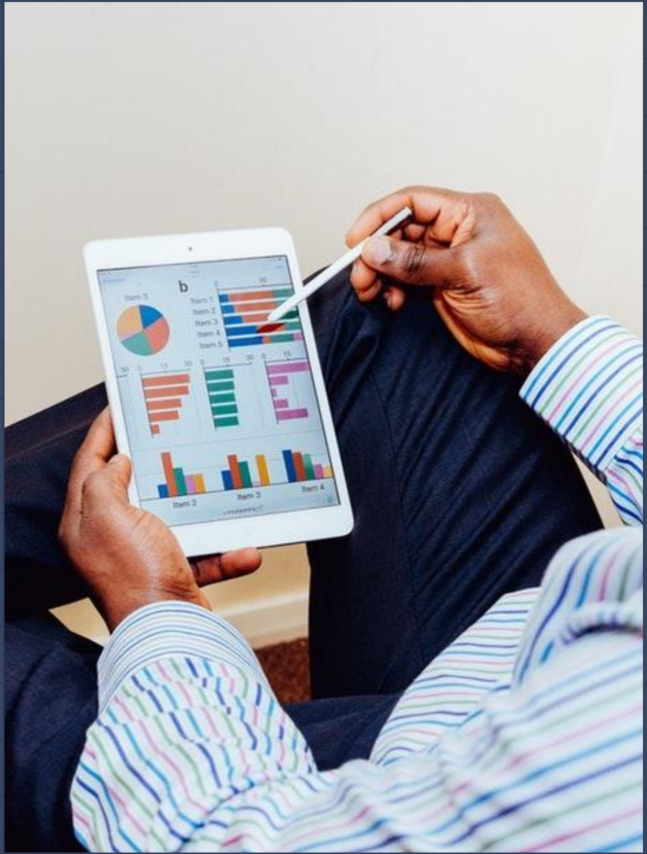
TRADING ACTION CLASSIFICATION

THU PHAM

FALL 2019

1. Introduction
2. Progress Report
3. Data Overview
4. Model Implementation
5. Report Poster



INTRODUCTION

- Utilize historical data, financial indicators, and news and social media analysis to apply machine learning methods.
- Forecast the movement of stock in a daily basis.
- Focus on trading action (long/short position).

PROGRESS REPORT

10/16/2019
Complete retrieved stock historical data, NYTimes data, and 7 days of Tweets. Complete sentimental analysis on News articles.

10/27/2019
Implementation of regression and classification models (Logistic Regression, SVM, Neural Networks)

11/15/2019
Conduct a new dataset using the existing one and tweet data obtain. Implement models to compare.

11/20/2019
Model performance evaluation and optimizations.

11/25/2019
Model performance evaluation and optimizations.

DATA OVERVIEW

Historical Stock Price

- Alpha Vantage API
- Daily
- Since 2000

	timestamp	open	high	low	close	volume
timestamp						
2019-10-15	2019-10-15	236.39	237.64	234.88	235.32	19012889
2019-10-14	2019-10-14	234.90	238.13	234.67	235.87	24106900
2019-10-11	2019-10-11	232.95	237.64	232.31	236.21	41698900
2019-10-10	2019-10-10	227.93	230.44	227.30	230.09	28253400
2019-10-09	2019-10-09	227.03	227.79	225.64	227.03	18692600
2019-10-08	2019-10-08	225.82	228.06	224.33	224.40	27955000
2019-10-07	2019-10-07	226.27	229.93	225.84	227.06	30576500
2019-10-04	2019-10-04	225.64	227.49	223.89	227.01	34619700
2019-10-03	2019-10-03	218.43	220.96	215.13	220.82	28606500
2019-10-02	2019-10-02	223.06	223.58	217.93	218.96	34612300

DATA OVERVIEW

Convert closing price to classification data

- Long stock (1) if the price increase by 2.5%.
- Short stock (0) if otherwise.

	timestamp	open	high	low	close	volume	prev_close	action
timestamp								
2019-10-15	2019-10-15	236.39	237.64	234.88	235.32	19012889	235.87	0.0
2019-10-14	2019-10-14	234.90	238.13	234.67	235.87	24106900	236.21	0.0
2019-10-11	2019-10-11	232.95	237.64	232.31	236.21	41698900	230.09	1.0
2019-10-10	2019-10-10	227.93	230.44	227.30	230.09	28253400	227.03	0.0
2019-10-09	2019-10-09	227.03	227.79	225.64	227.03	18692600	224.40	0.0
2019-10-08	2019-10-08	225.82	228.06	224.33	224.40	27955000	227.06	0.0
2019-10-07	2019-10-07	226.27	229.93	225.84	227.06	30576500	227.01	0.0
2019-10-04	2019-10-04	225.64	227.49	223.89	227.01	34619700	220.82	1.0
2019-10-03	2019-10-03	218.43	220.96	215.13	220.82	28606500	218.96	0.0
2019-10-02	2019-10-02	223.06	223.58	217.93	218.96	34612300	224.59	0.0

DATA OVERVIEW

New York Times API

- Headlines and articles archive from the past 20 years
- Merge to the stock historical data frame
- Sentimental analysis

	timestamp	open	high	low	close	volume	prev_close	action	neg	neu	pos
2000-01-03	2000-01-03	104.8750	112.5000	101.6880	111.938	133949200	102.813	1.0	0.051	0.871	0.078
2000-01-04	2000-01-04	108.2500	110.6250	101.1880	102.500	128094400	111.938	0.0	0.056	0.904	0.039
2000-01-05	2000-01-05	103.7500	110.5630	103.0000	104.000	194580400	102.500	0.0	0.093	0.828	0.079
2000-01-06	2000-01-06	106.1183	107.0000	95.0000	95.000	191993200	104.000	0.0	0.079	0.835	0.086
2000-01-07	2000-01-07	96.5000	101.0000	95.5000	99.500	115183600	95.000	1.0	0.072	0.838	0.090
2000-01-10	2000-01-10	102.0000	102.2500	94.7500	97.750	126266000	99.500	0.0	0.081	0.850	0.068
2000-01-11	2000-01-11	95.9380	99.3750	90.5000	92.750	110387200	97.750	0.0	0.086	0.846	0.069
2000-01-12	2000-01-12	95.0000	95.5012	86.5000	87.188	244017200	92.750	0.0	0.115	0.789	0.096
2000-01-13	2000-01-13	94.4840	98.7500	92.5000	96.750	258171200	87.188	1.0	0.097	0.818	0.085
2000-01-14	2000-01-14	100.0000	102.2500	99.3750	100.438	97594000	96.750	1.0	0.097	0.832	0.071

2,131,988

headlines

DATA OVERVIEW

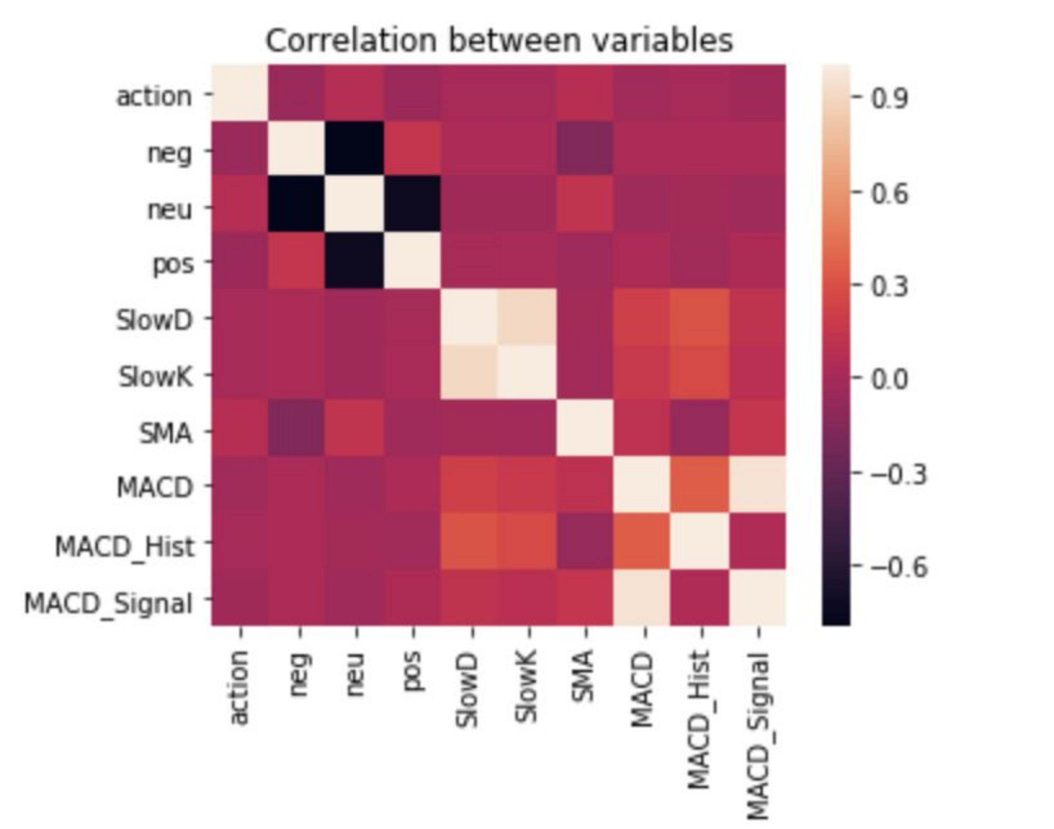
Twitter Static Sentimental Analysis from Kaggle

- AAPL
- 2016-2019
- Daily

date	ts_polarity	twitter_volume
1/1/16	0.11969256	417
1/2/16	0.14077416	495
1/3/16	0.18113164	518
1/4/16	0.07038878	1133
1/5/16	0.13363479	1430
1/6/16	0.07204194	1949
1/7/16	0.07436948	2289
1/8/16	0.05159477	2235
1/9/16	0.03234206	892
1/10/16	0.14592163	625
1/11/16	0.01944325	1222
1/12/16	0.12136357	1293

DATA OVERVIEW

	timestamp	open	high	low	close	volume	prev_close	action	neg	neu	pos	SlowD	SlowK	SMA	MACD	MACD_Hist	MA
0	2000-01-03	104.8750	112.500	101.688	111.938	133949200	102.813	1.0	0.051	0.871	0.078	86.9150	83.4605	227.842	4.9364	0.9771	
1	2000-01-04	108.2500	110.625	101.188	102.500	128094400	111.938	0.0	0.056	0.904	0.039	85.5056	89.1786	226.710	4.5060	0.7911	
2	2000-01-05	103.7500	110.563	103.000	104.000	194580400	102.500	0.0	0.093	0.828	0.079	80.6293	88.1060	225.310	4.0429	0.5257	
3	2000-01-06	106.1183	107.000	95.000	95.000	191993200	104.000	0.0	0.079	0.835	0.086	77.2606	79.2323	224.069	3.5916	0.2058	
4	2000-01-07	96.5000	101.000	95.500	99.500	115183600	95.000	1.0	0.072	0.838	0.090	74.7201	74.5496	223.276	3.4743	0.1400	

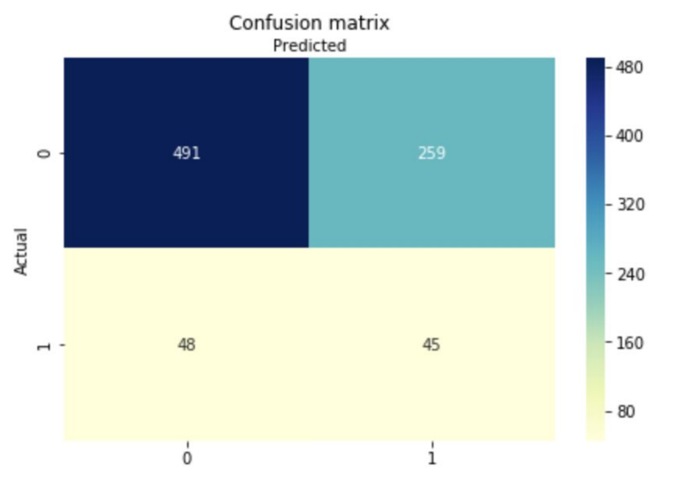

Correlation between variables
action
neg
neu
pos
SlowD
SlowK
SMA
MACD
MACD_Hist
MACD_Signal
0.9
0.6
0.3
0.0
−0.3
−0.6

DATA OVERVIEW

timestamp	action	neg	neu	pos	SlowD	SlowK	SMA	MACD	MACD_Hist	MACD_Signal	ts_polarity	twitter_volume
2016-01-22	1.0	0.109	0.809	0.082	79.6606	85.1923	75.0608	2.4191	0.1092	2.3099	0.096273	1200.0
2016-01-29	1.0	0.097	0.836	0.066	67.2965	65.6720	73.7313	2.4077	0.1548	2.2529	0.082394	1542.0
2016-02-16	1.0	0.060	0.843	0.097	81.0210	86.3498	68.1872	1.5980	0.0426	1.5553	0.041126	1272.0
2016-03-01	1.0	0.059	0.835	0.106	74.2727	63.1132	65.8090	1.7590	-0.0708	1.8298	0.042563	1427.0
2016-05-16	1.0	0.091	0.811	0.098	33.0112	18.7889	62.0915	-1.9516	-0.4589	-1.4927	0.077478	2587.0
2016-07-27	1.0	0.142	0.757	0.101	9.5907	10.9260	63.9175	-2.3736	-0.4521	-1.9215	0.069447	5173.0
2016-09-14	1.0	0.115	0.801	0.084	37.0065	14.7593	74.4610	-0.4996	-1.0110	0.5114	0.074549	3897.0
2016-09-15	1.0	0.100	0.815	0.086	54.0529	39.3077	74.8690	-0.2453	-1.0094	0.7641	0.111114	3610.0
2016-11-16	1.0	0.091	0.795	0.115	72.2544	53.7573	66.7390	3.4828	0.4052	3.0776	-0.007081	1366.0

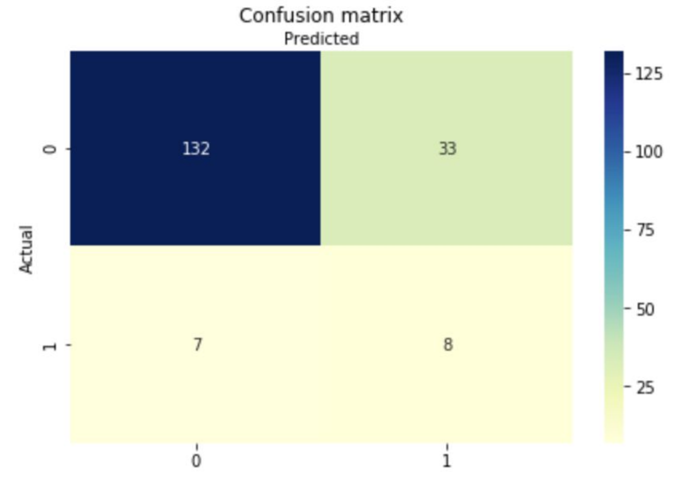
LOGISTIC REGRESSION

NY Times Sentimental



Accuracy: 63.58%

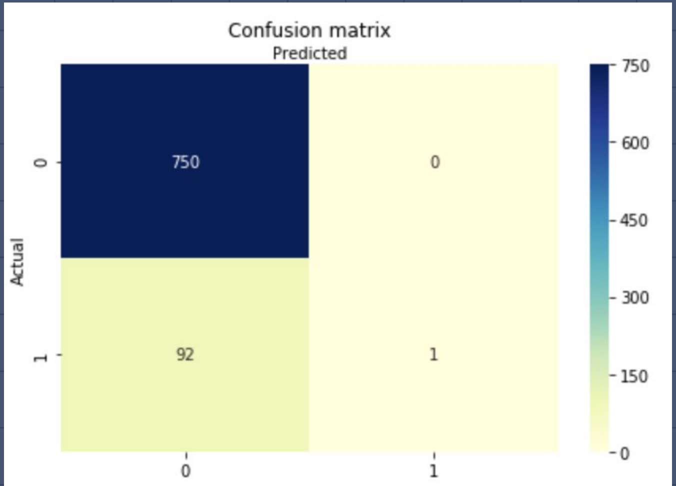
NY Times + Twitter Sentimental



Accuracy: 77.78%

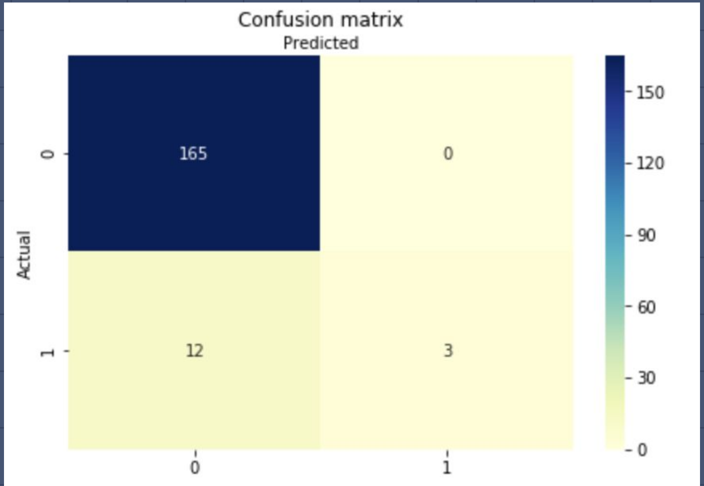
SUPPORT VECTOR MACHINE

NY Times Sentimental



Accuracy: 89.09%

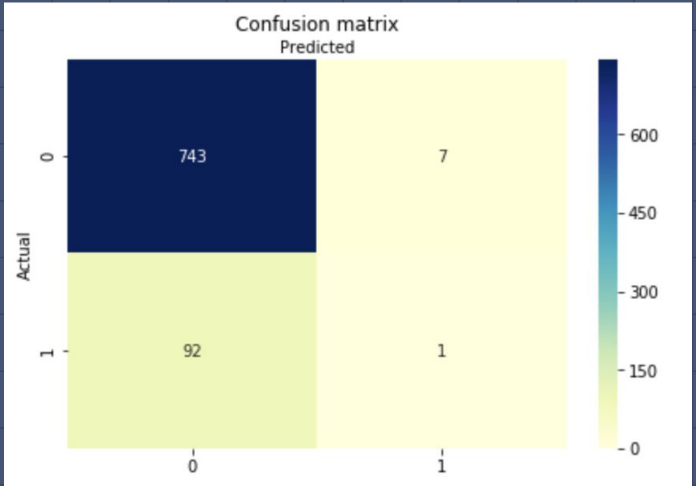
NY Times + Twitter Sentimental



Accuracy: 93.33%

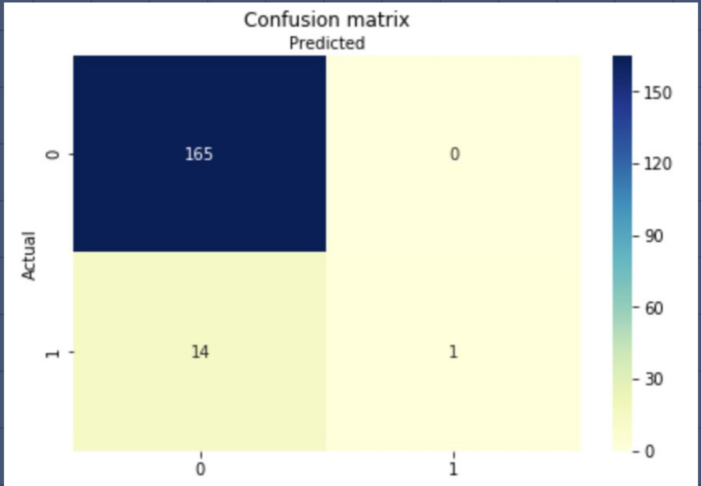
KERAS SEQUENTIAL

NY Times Sentimental



Accuracy: 88.26%

NY Times + Twitter Sentimental



Accuracy: 92.22%

Trading Action Classification using Machine Learning

Thu Pham

Luther College, Department of Computer Science & Data Science

ABSTRACT

Stock prices fluctuate within seconds and are affected by complicated financial and non-financial indicators. As opposed to predicting the trend in short-term which is used in the high-frequency trading market, our intention is to forecast the upward and downward movement in the weekly-basis not solely for algorithmic trading, but as a supplement to help investors alike on decision-making.

Our project is currently only applied for APPLE Stock.

DATA SOURCE

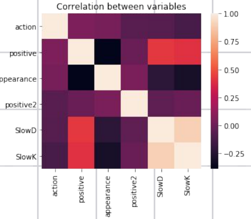
The project uses the free API from Alpha Vintage (alphavantage.co) to the monthly stock market price historical data in the past 20 years.

Additionally, Alpha Vintage also provides additional financial indicators, such as the STOCH index data, moving average (SMA) values, moving average convergence / divergence (MACD) values.

Most importantly, the project uses the New York Times Articles API to retrieve all the news headlines New York Times published since January 2000 and static Twitter data from 2006.

DATA PREPROCESSING

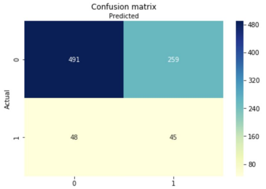
To preprocess the stock market price historical data, we set an expected return value (for example, 2.5%), which is minimal change in the stock price compared to the previous month for a long position. With this threshold value, we add a new variable called **action** with 2 values: 1 represents long position and 0 represents short position. Then we use sentimental analysis to analyze the New York Times headlines.



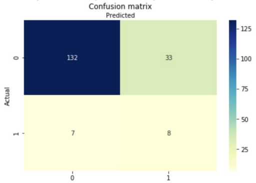
MODEL IMPLEMENTATION

Logistic Regression (LR)

Logistic regression is a simple linear model for classification. The confusion matrix for each dataset is presented below:



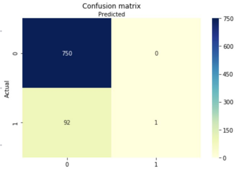
Without Twitter parameters: 63.58% accuracy



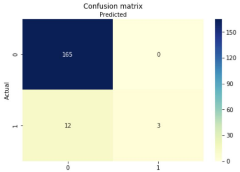
With Twitter parameters: 77.78% accuracy

Support Vector Machine (SVM)

Similar to Logistic Regression, SVM is an algorithm used for classification problems. However, in large dataset, SVM performs marginally better and less sensitive to outliers than Logistic Regression.



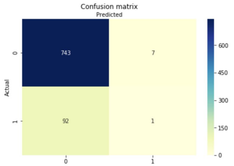
Without Twitter parameters: 89.09% accuracy



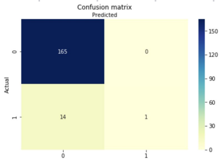
With Twitter parameters: 93.33% accuracy

Neural Networks

Since we are performing binary classification, a multi-layer perceptron is an appropriate method for such model. For the first dataset, we implement **LeakyReLU(alpha=0.5)** for one layer kernel. For the Twitter dataset, we implement a **Dense** layer, which is a connect layer; the first two layers take **activation** argument **tanh** while the last layer takes **stigmoid** .



Without Twitter parameters: 88.26% accuracy



With Twitter parameters: 92.22% accuracy

RESULT

The models successfully predict the actions at least 63% of the time with the expected return value close to 0. The expected return value significantly affects the accuracy of the models.

The smaller the expected return value, the more likely the decision making fluctuates with the market, so the more least accurately the models predict.

Generally, models predicting the later dataset with added Twitter parameters have higher accuracy than the original dataset without them. However, we need to take note that the Twitter dataset is much smaller than the original dataset.

CONCLUSION

With the results of this project, we can conclude that social media have an effect on the stock market since our dataset including Twitter parameters has higher accuracy in model training. By training our data in a daily basis, we have a bigger dataset to capture the volatile nature of stock.

Additionally, as we analyze the dataset using three different Machine Learning methods, including one using complex neutral networks, the results are improved but not significantly.

While we have a high accuracy for some models, we believe that we could do better with more data and time, especially if we have access to Twitter daily data.

ACKNOWLEDGEMENT

Thank you Professor Shafqat Shad of Data Science Department at Luther College for your support and instructions throughout this project.

The project is available at: github.com/thu2pham/StockForecasting

THANKS!

Any questions?

Find me at:

https://github.com/thu2pham/tradingclassification