Stock Market Trends Prediction after Earning Release

Stanford | ENGINEERING
Electrical Engineering

Chen Qian, Wenjie Zheng, Ran An
(cqian23 / zhwj814 / anran)@stanford.edu

Stanford University

Introduction

Public companies release quarterly earnings report to present financial status to investors. Short term stock prices can be significantly affected by company's quarterly performance. Our project aims to explore the impact of financial report factors, earning per share (EPS), consensus forecast and the earning report related articles on short term stock market trends. We narrow down our targeted company to Silicon based public technology companies.

Data Selection

- ***Financial factors***

Total asset	Total equity	Total revenue	Net income

- ***News and analytic articles***

We would like to analyze if these articles from mainstream media could influence stock market and investor's trading strategies.

- ***EPS vs Forecast***

How much the actual EPS bits consensus expectation is also a key factor to be considered for investment strategy.

- ***Training labels***

Label 1: Next day opening price vs release date close price
Label 2: Next day closing price vs release date close price

Data Preprocessing

Relativisation
Calculate the growth rate of each financial factor from last quarter, allows us to combine companies data together.

Sentiment Analysis
Interact with Stanford NLP toolkit to obtain sentiment analysis matrix from news and articles.

Normalization
Normalized input features simplifies model design.

Features

Relativized financial data

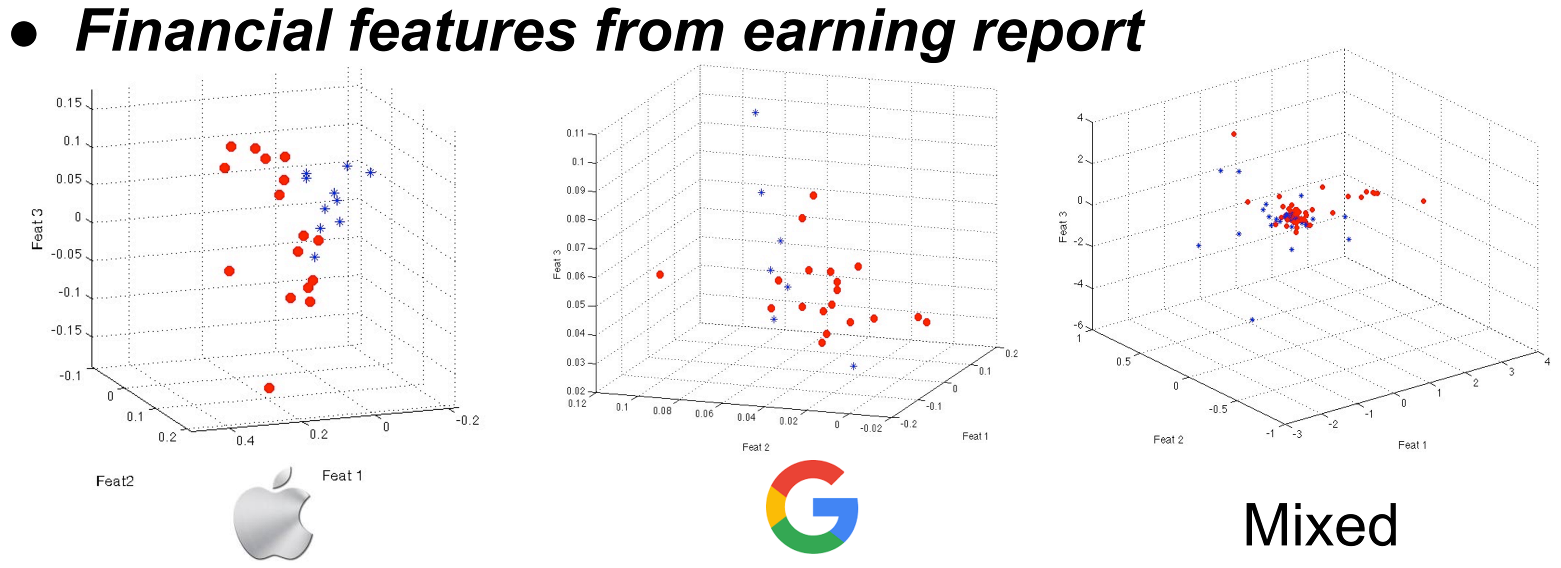
EPS surprise	Total asset	Total equity	Total revenue	Net income
15.30%	5.52%	7.82%	-3.45%	14.21%

Normalized sentiment analysis features

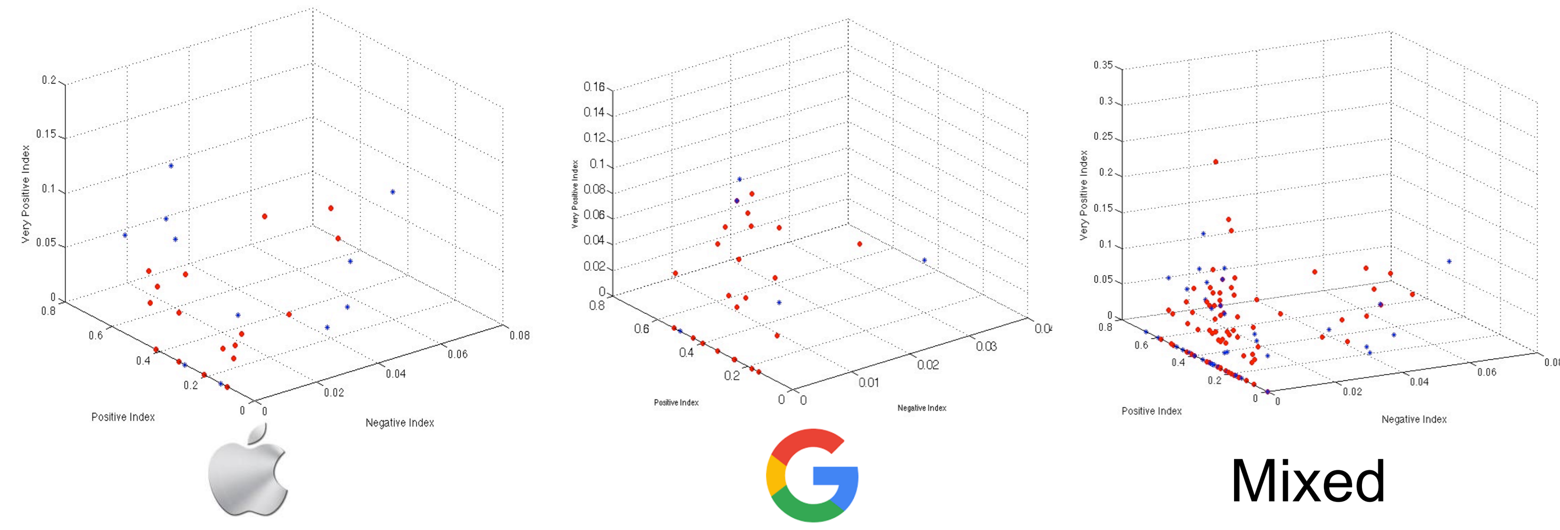
Very negative	Negative	Neutral	Positive	Very Positive
2.564%	8.974%	62.820%	20.512%	5.128%

Data Visualization

- ***Financial features from earning report***



- ***Sentiment features from articles***



Model Selection & Evaluation

Model	Equation	Parameters	Smoothing Tools
Locally Weighted LR	$\beta = (X'WX)^{-1}X'WY$	Bandwidth T	$D = ae^{-\left(\frac{\|X-X0\|}{2T}\right)}$

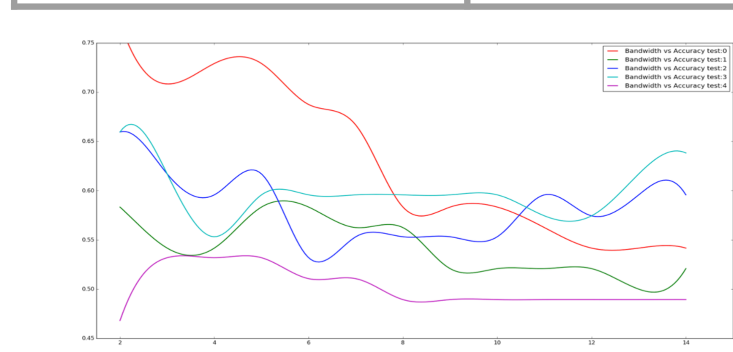


Figure. LWLR with financial features: Bandwidth vs Accuracy

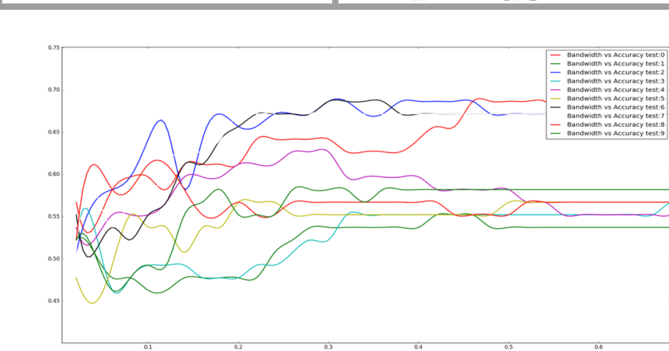


Figure. LWLR with NLP features: Bandwidth vs Accuracy

Model	Equation	Parameters	Kernel
SVM	$J_\lambda(\alpha) = \frac{1}{m}\sum_{i=1}^{m} L(K^{(i)T}\alpha, y^{(i)}) + \frac{\lambda}{2}\alpha^T K\alpha$	Bandwidth Γ	$K(x,x') = ae^{-\left(\frac{\|x-x'\|}{2\Gamma^2}\right)}$

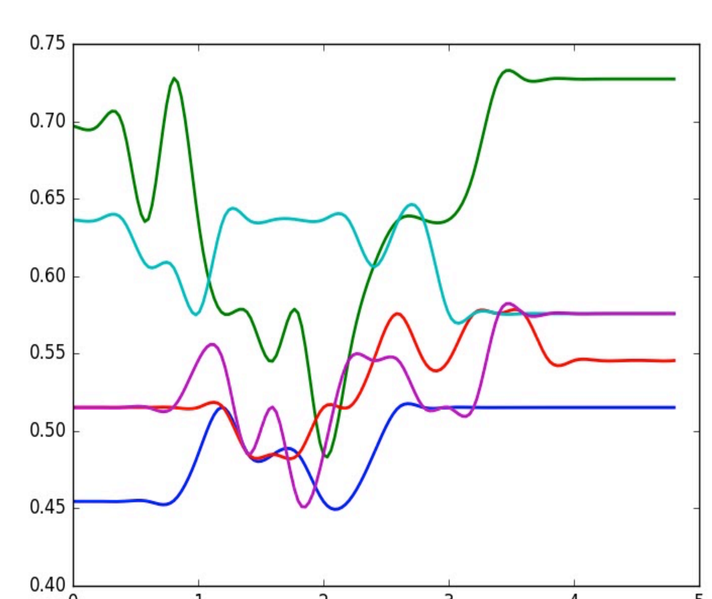


Figure. SVM Kernel Gamma Γ vs Accuracy

1. In SVM with RBF, decision boundary to separate samples of different labels is easier to find with a certain company. It is hard to find the boundary if we mix data among companies. However, in terms of article, no matter if the focus is on individual company or several, the data tends to follow a similar distribution.
2. In experiments with SVM and LWLR, we tried to adjust the bandwidth τ/Γ to improve the performance of these models.
3. 5-fold cross validation was exploited for accuracy measurement

Results

- **All error analysis are validated using K-Fold cross validation (K = 5)**

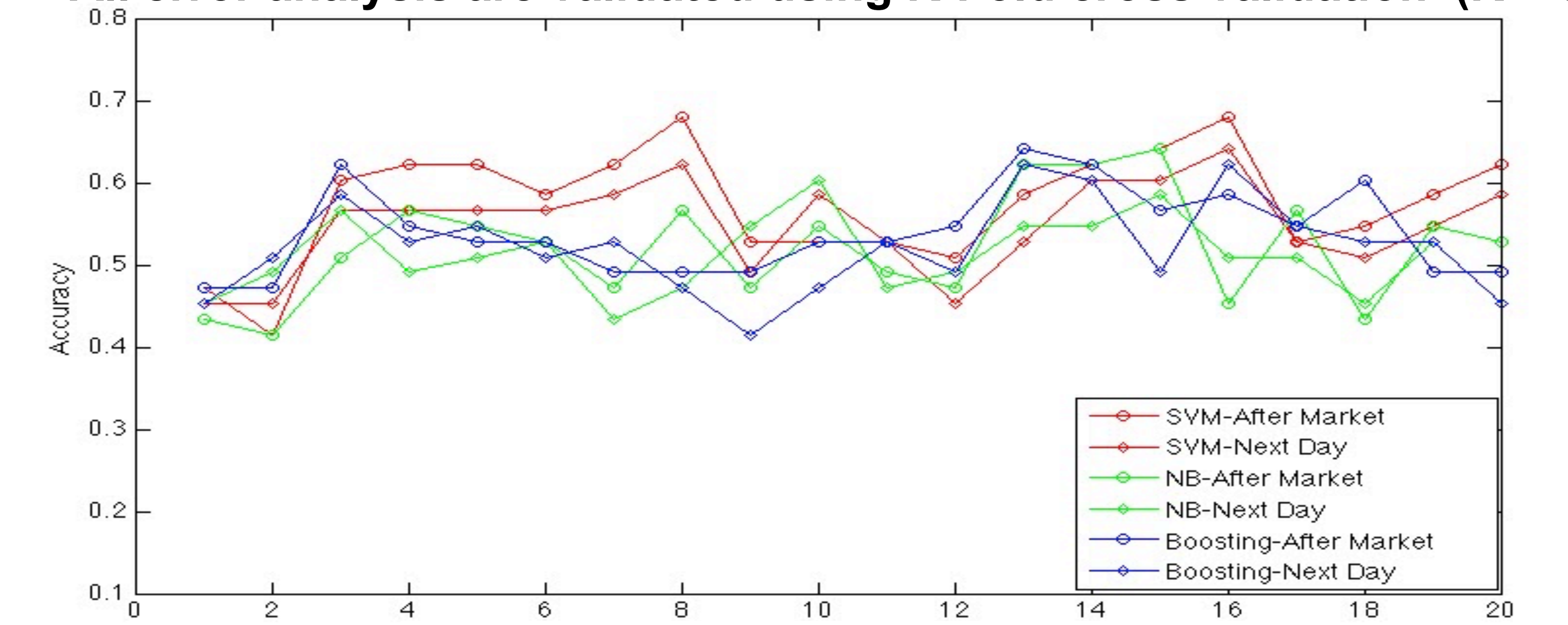


Figure. Accuracy with different models over multiple tests

	Classification Models			
Predict by report	**SVM**	**Naïve Bayesian**	**Locally Weighted**	**Boosting**
after-market	**69.80%**	**66%**	**64.15%**	**66%**
next-day	**64.20%**	**60.30%**	**68.12%**	**58.30%**
Predict by article (norm)				
after-market	**57.64%**	**52.17%**	**58.96%**	**53.96%**
next-day	**54.91%**	**50.94%**	**64.38%**	**52.17%**

Discussion

Result: Stock market is known as a chaotic system and our model analysis depicts that even building our model with empirical key features could still result in low accuracy. However, by limiting our scope to earning release day, we are able to build the prediction model of around 70% accuracy.

Models and features: SVM seems to give us a better prediction than NB and other generalized LR algorithms. Further, it is observed that data visualization from different companies shows that data set from a certain company is much more distinguishable than mixing data together. The commercial correlations between companies are far less than previous expectation and the features we collected are far from enough to predict the stock trend.

Limitation: Due to the limited number of company choices, we have small data size (~300 samples) which could lead to high bias and over-fitting. Stock price is not only affected by certain financial features, consensus news, but also company direction and future business guidance, which are difficult to be digitized.

Future

Feature selections: Write script to auto collect/process data to increase the learning set size and use feature ranking to analysis the covariance between features and labels and pick the most top ranked one as the model inputs.

NLP analysis: The sentiment result from Stanford NLP toolkit is too general to apply for financial news analysis and we would spend more time to develop a financial specific sentiment analysis algorithm.

Model Improvement: We would spend more time on SVM algorithm and explore the most appropriate kernel to separate features in high dimension.