Machine Learning Engineer Nanodegree

Capstone Project Proposal

“Greedy Fear”
Predicting Stock Markets using Deep Learning NLP to Analyze Financial Statements

Ming-Hao Yu

March 1, 2020

1 Domain Background

"Be fearful when others are greedy and greedy when others are fearful." - Warren Buffett.

Warren Buffett is one of the greatest investors in the world. He has been my ideal role model since the first time I read his biograph in my childhood. His investment philosophy, value investing, emphases investing equities that are underpriced than their intrinsic values, which aspires my perspective toward investment.

Nowadays, various financial service companies such as Robinhood, Charles Schwab and TD Ameritrade offer zero commission fee and fraction share transaction that makes trading available and becoming a hobby to a tremendous population of common people in the United States. However, inexperienced investors are not rational and there are researches [1][2] in behavioral finance empirically evidence that those people tend to overreact to new information.

To succeed in stock markets, we not only need to avoid overreacting to the news but also would like to exploit people's overreaction. When people are dumping their shares due to the fear of some pessimistic news released, we are going to collect the underpriced equities as getting a discount from the stock markets just as Buffett's wisdom. Therefore, the project – "Greedy Fear" aims to utilize the NLP technique and Machine Learning models to analyze financial news and to predict the stock prices to keep us calm and cool when people are in fear and greedy.

2 Problem Statement

The main objective of the project is to utilize Natural Language Processing (NLP) and Deep Learning techniques to analyze companies' 8-K Forms [3], an unscheduled report required to announce major business events relevant to shareholders, and to generate signals for stock price performance. We will predict the stock price direction (appreciate or depreciate) on the next day the 8-K Forms released as a classification problem.

3 Datasets and Inputs

The datasets consist of two parts, financial statements and stock price data. The 8-K Forms are available publicly in the SEC EDGAR database. The historical stock price data are available on Yahoo Finance. Both the two datasets can be collected through web scraping.

Data Description:

1. Stock price data (Float): Historical Open, High, Low, Close price of companies.

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Return
2020-02-14	337.51	337.73	336.20	337.60	64582200	0.0	0	0.001600
2020-02-18	336.51	337.67	335.21	336.73	57226200	0.0	0	-0.002584
2020-02-19	337.79	339.08	337.48	338.34	48814700	0.0	0	0.004759
2020-02-20	337.74	338.64	333.68	336.95	74163400	0.0	0	-0.004125
2020-02-21	335.47	335.81	332.58	333.48	113788200	0.0	0	-0.010405
2020-02-24	323.14	333.56	321.24	322.42	161088400	0.0	0	-0.034303
2020-02-25	323.94	324.61	311.69	312.65	218913200	0.0	0	-0.031249

(stock price data)

2. 8-K Forms (Textual): Historical finance announce of companies, industry of the business and date.

	ticker	cik	GICS Sector	text	release_date
0	A	1090872.0	Health Care	0001564590-18-006570.txt : 20180322 0001564590...	2018-03-22 16:22:07
1	A	1090872.0	Health Care	0001090872-18-000002.txt : 20180214 0001090872...	2018-02-14 16:27:02
2	A	1090872.0	Health Care	0001564590-18-000605.txt : 20180118 0001564590...	2018-01-18 16:09:52
3	A	1090872.0	Health Care	0001090872-17-000015.txt : 20171120 0001090872...	2017-11-20 16:09:02
4	A	1090872.0	Health Care	0001090872-17-000011.txt : 20170815 0001090872...	2017-08-15 16:12:29
5	A	1090872.0	Health Care	0001564590-17-013595.txt : 20170719 0001564590...	2017-07-18 18:38:27
6	A	1090872.0	Health Care	0001564590-17-013541.txt : 20170717 0001564590...	2017-07-17 16:06:17
7	A	1090872.0	Health Care	0001090872-17-000006.txt : 20170522 0001090872...	2017-05-22 16:09:54
8	A	1090872.0	Health Care	0001564590-17-004619.txt : 20170316 0001564590...	2017-03-16 16:06:36
9	A	1090872.0	Health Care	0001090872-17-000002.txt : 20170214 0001090872...	2017-02-14 16:13:30

(8-K Forms textual data)

```
print(df['text'][3])
```

0001090872-17-000015.txt : 20171120 0001090872-17-000015.hdr.sgml : 20171120 20171120160902 ACCESSION NUMBER: 0001090872-17-000015 CONFORMED SUBMISSION TYPE: 8-K PUBLIC DOCUMENT COUNT: 2 CONFORMED PERIOD OF REPORT: 20171120 ITEM INFORMATION: Results of Operations and Financial Condition ITEM INFORMATION: Financial Statements and Exhibits FILED AS OF DATE: 20171120 DATE AS OF CHANGE: 20171120 FILER: COMPANY DATA: COMPANY CONFORMED NAME: AGILENT TECHNOLOGIES INC CENTRAL INDEX KEY: 0001090872 STANDARD INDUSTRIAL CLASSIFICATION: LABORATORY ANALYTICAL INSTRUMENTS [3826] IRS NUMBER: 770518772 STATE OF INCORPORATION: DE FISCAL YEAR END: 1031 FILING VALUES: FORM TYPE: 8-K SEC ACT: 1934 Act SEC FILE NUMBER: 001-15405 FILM NUMBER: 171213843 BUSINESS ADDRESS: STREET 1: 5301 STEVENS CREEK BLVD CITY: SANTA CLARA STATE: CA ZIP: 95051 BUSINESS PHONE: (408) 345-8886 MAIL ADDRESS: STREET 1: 5301 STEVENS CREEK BLVD, MS 1A-LC STREET 2: P.O. BOX 58059 CITY: SANTA CLARA STATE: CA ZIP: 95052-8059 FORMER COMPANY: FORMER CONFORMED NAME: HP MEASUREMENT INC DATE OF NAME CHANGE: 19990716 8-K 1 form8-kxq417pressrelease.htm 8-K Document UNITED STATESSECURITIES AND EXCHANGE COMMISSIONWashington, D.C. 20549 FORM 8-K CURRENT REPORTPursuant to Section 13 or 15(d) ofThe Securities Exchange Act of 1934 Date of Report (Date of earliest event reported): November 20, 2017 AGILENT TECHNOLOGIES, INC.(Exact name of registrant as specified in its charter) Registrant’s telephone number, including area code (408) 345-8886 (Former name or former address, if changed since last report.) Check the appropriate box below if the Form 8-K filing is intended to simultaneously satisfy the filing obligation of the registrant under any of the following provisions: o Written communications pursuant to Rule 425 under the Securities Act (17 CFR 230.425) o Soliciting material pursuant to Rule 14a-12 under the Exchange Act (17 CFR 240.14a-12) o Pre-commencement communications pursuant to Rule 14d-2(b) under the Exchange Act (17 CFR 240.14d-2(b)) o Pre-commencement communications pursuant to Rule 13e-4(c) under the Exchange Act (17 CFR 240.13e-4(c))Indicate by check mark whether the registrant is an emerging growth company as defined in as defined in Rule 405 of the Securities Act of 1933 (§230.405 of this chapter) or Rule 12b-2 of the Securities Exchange Act of 1934 (§240.12b-2 of this chapter).Emerging growth company o If an emerging growth company, indicate by check mark if the registrant has elected not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange Act. o Item 2.02. Results of Operations and Financial Condition. The information in this Item 2.02 of Form 8-K and Exhibit 99.1 attached hereto is furnished and shall not be deemed “filed” for purposes of Section 18 of the Securities Exchange Act of 1934, as amended, nor shall it be deemed incorporated by reference in any filing under the Securities Act of 1933, as amended. On November 20, 2017, Agilent Technologies, Inc. (the “Company”) issued its press release announcing financial results for the fourth fiscal quarter ended October 31, 2017. A copy of this press release is attached as Exhibit 99.1. We provide non-GAAP financial information in order to provide meaningful supplemental information regarding our operational performance and to enhance our investors’ overall understanding of our core current financial performance and our prospects for the future. We believe that our investors benefit from seeing

(Sample of 8-K forms)

Label:

1. Stock price direction (Integer): [1, 0, -1] represents [Increase, Neutral, Decrease] in stock price. Half of the labels are 0 (percentage change less than 0.5%), and the rest of the two labels distribute evenly.

4 Solution Statement

Both Shallow (non-Deep) and Deep NLP techniques would be applied to the textual data. TF-IDF (Term Frequency – Inverse Document Frequency) suits for extracting keywords from a document. To capture the semantic information from a document, we will utilize the state-of-the-art Transfer Learning model – BERT (Bidirectional Encoder Representations from Transformer) to generate numerical vectors (features) through the Pre-Training model and then Fine-Tune our ML model for our classification problem.

Besides the textual data, stock price trends in the market is the other key factor in trading. Thus, we would incorporate stock prices (OHLC, Open-High-Low-Close), returns and some investment indicators of previous stock market day in order to capture the market momentums.

After combining features engineered from both textual and numerical inputs, we will build a deep neural network model to perform the classification task.

5 Benchmark Model

The benchmark model would be common classification models such as Logistic Regression, Random Forest Classifier and XGBoosting Classifier. I will compare the performance of the neural network model and the above Shallow Learning Classifier.

6 Evaluation Metrics

We will define the stock market direction into three categories:

(1) Appreciation (Label: 1): price raises > 0.5%
(2) Depreciation (Label: -1): price decreases > 0.5%
(3) Neutral (Label: 0): price changes <= 0.5%

For the multi-class classification problem, we will use Accuracy as the ultimate metric to evaluate the model performance:

- Expected = [1, 1, 0, -1, -1, 0]
- Predicted = [0, 1, 1, 1, -1, 0]
- Correct Prediction = [“False”, “True”, “False”, “False”, “True”, “True”]
- Accuracy Score is $\frac{3}{6} = 0.5$

7 Project Design

- Programming Language:
  - Python 3.6
- Packages:
  - Pandas
  - Numpy
  - Matplotlib
  - Sklearn
  - TensorFlow PyTorch
  - Requests
  - NLTK
  - Beautiful Soup
  - yfinance
- Workflow

(1) Data Preparation
Collect stock price data from Yahoo Finance and SEC EDGAR database.

(2) Data Preprocessing
Resample the data to get a balanced dataset with each label evenly distributed. Perform Feature Engineering to generate word embedding features and numerical indicators.

(3) Train-Test Split
Split the data into 60-20-20 as training, validation and testing set.

(4) Model Training and Evaluation
We will first construct our Deep Learning neural network model by using PyTorch. Then we will start to build other benchmark models using Sklearn to compare the performance.

8 Reference

[1] Collin-Dufresne, Pierre and Johannes, Michael and Lochstoer, Lars A., “Asset Pricing When 'This Time is Different'” (January 15, 2016). Swiss Finance Institute Research Paper No. 13-73; Columbia Business School Research Paper No. 14-8. (http://dx.doi.org/10.2139/ssrn.2373495)

[2] Reza S. Mahani, Allen M. Poteshman, “Overreaction to stock market news and misevaluation of stock prices by unsophisticated investors: Evidence from the option market”. Journal of Empirical Finance, Volume 15, Issue 4, 2008 (https://doi.org/10.1016/j.jempfin.2007.11.001)

[3] Yusuf Aktan, “Using NLP and Deep Learning to Predict Stock Price Movements” (https://towardsdatascience.com/using-nlp-and-deep-learning-to-predict-the-stock-market-64eb9229e102)