AutoPhrase for Financial Documents Interpretation

Shaoqing Yi, Zachary Ling, Joey Hou

February 2021

1 Introduction

Stock market is one of the most popular markets that the investors like to put their money in. There are millions of investors who participate in the stock market investment directly or indirectly, such as by mutual fund, defined-benefit plan. Certainly, there are many people who research on the stock market, and they all know the information takes an important role in the decision making. According to the Strong Form Market Efficiency Theorem, the stock price is only determined by the new information; otherwise, it will be a random walk. For example, if a new annual report claims the 50 percent increase in earning per share, the investors will expect the stock price to increase by 50 percent correspondingly. However, there are thousands of information in this market everyday, and the investors can only pay attention to few of them. Therefore, the investors, especially the individual investors without the help of professional financial analyst, can only get the parts of the whole information, so his investment decisions may be biased. In this project, we want to solve this real-world problem to apply the Phrase mining technique to forecast the stock price and help the investors to make the decision. By the inspiration of the AutoPhrase NLP model from Professor Shang[1], we are going to apply the AutoPhrase model to extract the high-quality phrases from the 8-K report, which described the big event happened in the company, and use machine learning to predict the stock price. We will also give some case studies to apply our model in the real financial investment. We hope our project can help the investors to make good decisions and get extra profit of the portfolio.

2 Literature Review

The researches of stock price prediction were a hot topic hundreds years ago, as long as the stock data were available to the investors. The investors were trying to find a solid method to predict the future price from the past trading data, such as price and volume, which was called the technical analysis. For example, in 1930s, the professional accountant Ralph Nelson Elliott discovered some special

price pattern, and he introduced the Elliott Wave Theorem to explain each of the patterns. Moreover, some investors also want to predict the stock performance based on the past operating data, such as profit margin, Earning per share. This is called the fundamental analysis. However, Our project about 8-K reports is a different analysis, called event-driven analysis. The event-driven analysis is a particular type of analysis based on the new information, especially the major events happened to the companies. These events, no matter good or bad, are the motivation of the price trend, because they can significantly change the expectation of the investors. The event-driven analysis is different from the previous technical analysis, because it is not based on the past information. It researches on the latest information and give the investment recommendation. Our project is a kind of event-driven analysis, which is to apply AutoPhrase model to extract the key information from the financial news and make the decision. There are many former event-driven analysis models. Heeyoung Lee and other three scholars apply unigram model to the Form 8-K to get the feature vector and train a linear model to predict the stock price.[6] However, in our project, we do the phrase mining with AutoPhrase model on the 8-K reports database and extract the high-quality phrases. We then get our feature vectors based on the apparent of the high-quality phrases on these 8-K reports, instead of unigram model. We want to do an empirical testing on the AutoPhrase to analyze whether the AutoPhrase model can help to extract a better phrase vector for stock price prediction.

3 Methodology

3.1 Overview

In this project, we will train a machine learning model based on the phrase vectors from the AutoPhrase model and compare with two baseline models. We are doing a classification task, instead of a regression task on the stock price. We label the price data as three price trend classes – "Up", "Stay" and "Down", based on the certain intervals of price change percentage. The first baseline model is the EPS model, which only uses the "Surpriseness" of the Earning per share to predict the price trend. The definition of "Surpriseness" will be described in the following section. The second baseline model is the linear model based on unigram given by Heeyoung Lee. He got around 50% accuracy on the testset, based only on unigram feature vectors. For this project, we will train our own models based on the AutoPhrase high-quality phrases and compare with the baseline models. We will also tune the hyper-parameters and compare different machine learning models, such as Random Forest, Logistic regression, SVM.

3.2 Data Source

We collect the data from two different data sources. The first one is the IFind Financial Data Terminal. We have collected about 2000 8-K report and corresponding price data from the terminal. In addition, we also have a 8-K DataBase (Size: 1.4G Click to Download)[6] from Stanford by Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky.

3.3 Price trend labels

For the stock price, we have the daily stock price for the 10 years. We calculate the overnight, 7-day, 14-day and 30-day price change, which is percentage change of a certain future day to the original day that the 8-K reports released. We also eliminate the effects of the systematic stock market change by subtracting the SP 500 stock index percentage change. If the price change percentage is higher than 1%, we label it as "Up". If the price change is between -1% and 1%, we label it as "Stay". If the price change is below -1%, we label it as "Down".

3.4 "Surpriseness" of EPS

We have the data for two kinds of EPS, the reported EPS and the concensus EPS, which is the estimated EPS provided by the stock researchers. The percentage difference between these two types of EPS is called the Surprise. If the reported EPS is higher than the concensus EPS, there is a positive Surprise, and we expect that the stock price will go up. If the reported EPS is lower than the concensus EPS, we expect the stock price will go down.

3.5 AutoPhrase

First, we apply the AutoPhrase model to our 8-K reports with the knowledge base quality terms from the Wikipedia provided by Professor Shang. We do some data analysis and visualization, but the outcome is not what we expect. There are many high-quality terms provided by AutoPhrase model, that are meaningless to a financial report. Based on the advice from Professor Shang, we find our own financial knowledge base. We do the web mining to the Investopedia website, which is a website to help people to study finance. We get about 7000 Financial Terms from the Investopedia and replace the wiki terms to the financial terms.

3.6 Models

We selected 3 different classifiers, each trained on our 3 different feature sets: baseline, baseline + unigram features, baseline + phrase features. Out of the Logistic Regression, SVM, and Random Forest classifiers, the Random Forest had the best results on the training, validation, and test set.

3.6.1 Baseline

Includes only financial features (earnings surprise, price movements from 7 to 365 days, volatility index) and event features. After tuning hyperparameters on the validation set, the best model was the Random Forest classifier with parameters, max depth = 10 and n estimators = 2000. This model gave a 51.94% on the test set.

3.6.2 Unigram + Phrase

To see if text features could add any additional value to our model, we utilized the baseline features along with either phrase or unigram features. Each model utilized 2107 text features, encoded as a vector with binary entries. After tuning hyperparameters, the final model included parameters, max depth = 10, n estimators = 2000, and max features = 1250, for both the unigram and the phrase model. These models generated accuracies of 52.56% and 52.61% respectively.

4 Exploratory Data Analysis

4.1 Dataset Overview

	num of 8-K's	num of words	num of firms
Train	17098	313867921	1410
Val	8720	164041583	1372
Test	9076	163871871	1380

Table 1: Base descriptive info of 8-K reports

The Table 1 shows the number of 8-K's, number of words, and number of companies per split (training, validation, and test). Despite the large number of words and 8-K's in our training set, the balance of firms (1444 total) throughout all of our splits can help our model have more balanced predictions across various firms.

Table 2 lists some of the most common event types (reason for filing an 8-K) within the training set. Since different events can drastically change the contents of an 8-K form, we thought it as a significant feature in identifying variance among groups.

Table 3 shows that our data consists of around 38% "down", 22% "stay", and 40% "up" labels. This breakdown is also roughly consistent within each split: train, validation, and test. This will allow for our training data to match the rest of the splits as best as possible. Though the "stay" labels only make up a small minority of the data, it is more important to better predict "down" and "up" due to its larger price swings.

event	count
financial statements and exhibits	32071
results of operations and financial condition	31271
regulation fd disclosure	4994
other events	2305
election of directors	923
entry into a material definitive agreement	550
appointment of certain officers	528
departure of directors or certain officers	528
appointment of principal officers	395
departure of directors or principal officers	395

Table 2: Event Type Frequency

target	DOWN	STAY	UP
all_date	38.13%	21.65%	40.22%
train	37.22%	21.38%	41.40%
val	39.24%	21.55%	39.20%
test	38.83%	22.29%	38.87%

Table 3: Label Breakdown

Phrase	% of Documents with Phrases
press release	0.8875
cash flow	0.5115
fourth quarter	0.5056
annual report	0.4219
income tax	0.3828
accounting principle	0.3780
accounting principle	0.3644
financial measures	0.3437
risk factor	0.3149
tax rate	0.2896

Table 4: Top 10 phrase Frequency

The Table 4 shows the distribution of the top 10 unigrams and phrases within our 8-K forms. The top phrases are seen in a smaller percentage of 8-K's compared to our unigrams which are all almost 90% and above in frequency. The high frequencies among unigrams may make it a less valuable feature because of the homogeneous depiction of 8-K's. The phrases on the other hand may be able to better reveal differences within different 8-K forms.

4.2 Tokenization and Documents Analysis

We tokenize our documents and label the documents based on the price change compared to the benchmark stock index. We analyze about the class distribution and sentence information. From the Figure 1, Classes are pretty balanced balanced: 60.4% outperformed the index and 39.6% underperformed compared to the index. From the Figure 2, Distribution of token frequencies and document length is essentially the same among both classes.

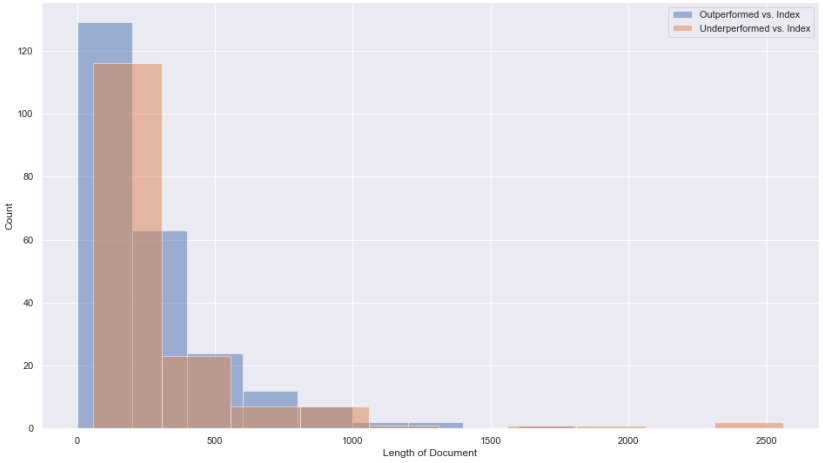

Figure 1: Length of documents

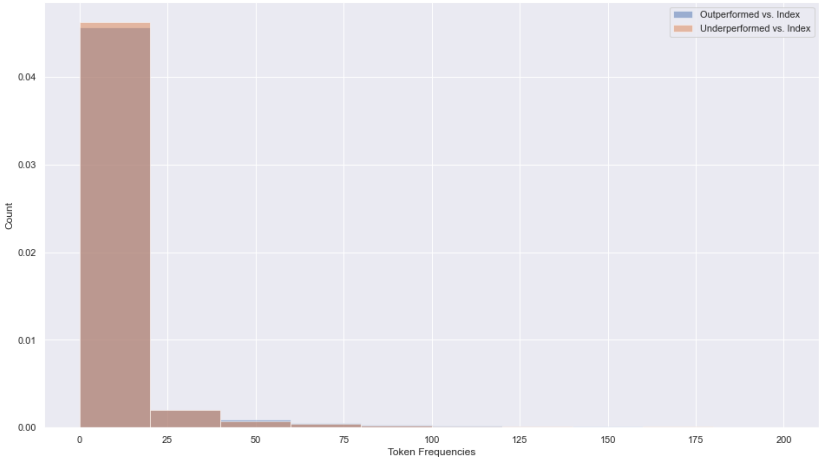

Figure 2: Token Frequency

4.3 Sentiment Analysis

We also conducted some sentiment analysis on the sentences in the 8-K reports. From Figure 3, we found out that the distribution of median subjectivity is different when comparing outperforming and underperforming securities.

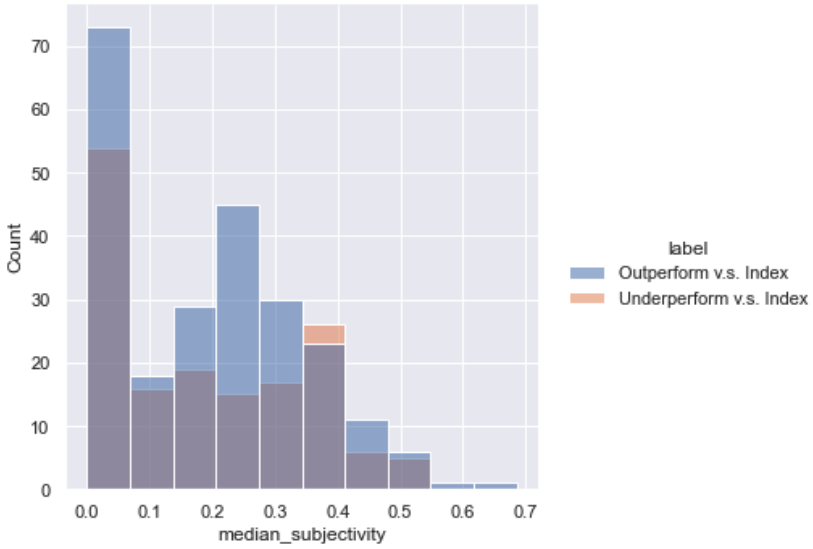

Figure 3: Median Subjectivity

4.4 AutoPhrase Result

Wiki Base	**Investopedia Base**
shuffle master	personal property
estee lauder	credit rating
stifel nicolaus	environmental liabilities
herman miller	contractual obligations
navigant consulting	severance benefits
sioux falls	accouting policies
teco coal's	annual salary
calvin klein	withholding tax
analog devices	service provider
novatel wireless	debt financing

Table 5: Top 10 quality phrases in Wiki and Investopedia Base

For Table 5, when utilizing the Wikipedia knowledge base, many of the resulting quality phrases were names of companies and people which does not

effectively depict the content of many 8-K forms. Therefore, through changing the knowledge base (pulled from Investopedia) we were able to output phrases that have semantic value within an 8-K form and the financial world.

Phrase	**% of Documents with Phrases**
financial condition	0.951866
financial statement	0.946368
financial statements	0.945491
press release	0.895192
financial result	0.725056
net income	0.711604
quarter end	0.657562
executive officer	0.634987
quarter ended	0.633641
chief executive	0.627208

Table 6: Top 10 phrase and quality in Investopedia Base

Unigram	**% of Documents with Phrases**
2	1
1	0.999649
financial	0.994034
results	0.97953
quarter	0.973681
company	0.972044
year	0.952451
net	0.944262
million	0.940578
income	0.886946

Table 7: Tope 10 unigram

The above shows the distribution of the top 10 multi-phrases (Table 6) and unigrams (Table 7) in our 8-K forms. The top phrases are seen in a smaller percentage of 8-K's compared to our unigrams which are all almost 90% and above in frequency. The high frequencies among unigrams may make it a less valuable feature because of the homogeneous depiction of 8-K's. The phrases on the other hand may be able to better reveal differences within different 8-K forms.

	Baseline	Unigram	Phrase
UP	51.64%	51.89%	51.68%
STAY	29.43%	43.15%	47.58%
DOWN	53.91%	54.14%	54.25%

Table 8: Class Accuracy

5 Result Analysis

5.1 Model Performance

Table 8 show's each model's accuracy broken down by labels. All of the models were able to predict "up" and "down" better than chance which has the potential to provide great benefits in the financial world. Though all three models perform relatively the same for the "up" and "down" label, the unigram and the phrase models are able to predict the "stay" label with much higher accuracy. For our test set, the phrase model had the highest accuracy among all the 3 labels but only by a very small margin.

Through using micro-averages and the OneVsRestClassifier, Figure 7 shows the estimated ROC curve for all three models. All three models are able to perform relatively the same in terms of sensitivity and specificity.

As shown by Figure 5 and 6, the model's most important features were dominated by the main numerical features: price changes and earnings surprise (the most dominant feature). The high predicting power of these features helps to explain the baseline's similar performance to the other enhanced models. Nevertheless, it is interesting to note some of the phrases and unigrams that contributed to the model's predictability such as "weak" or "revenue growth". Since many of these words/phrases make sense in a financial context, it helps to explain how some were able to have a small impact on the model, while words or phrases such as "gentleman" or "fee letters" have no impact on the model.

5.2 Simulation

We do a simulation to invest the stock market with our AutoPhrase model. We first train the models using data from 2002-2009. In each month of 2010-2012, we buy the stock with "UP" prediction result and short the stock with "DOWN" prediction, and calculate the average rate of return, assuming there is no commission fee. Note that this is the best way to simulation because the date and event of the 8-K reports are unpredictable. It is hard to buy all the stocks predicted to "UP", because we don't know the exact number of these stocks, and we are hard to assign the money to each of the stocks. In addition, the money needed for "short" is not same as "buy". Therefore, our simulation is not the best simulation based on the reality.

Figure 7 is the rate of return curve. (All the rate of reutrn is normalized by SP 500 Index.) In this chart, the three models perform very similar. The rate of return for AutoPhrase model is highest, which is 201.8%, and the EPS Baseline is the lowest, which is 197.1%. The return for Unigram model is in the middle, which is 201.2%.

6 Conclusion

From the paper, we can see that the linguistics factors, such as Unigram and Phrase mining, can help the prediction of the future price trend. However, the model with AutoPhrase does not outperforms the model with single unigram. Nonetheless, it is still a very interesting topic to apply the phrase mining to company news and help to make investment decision. We are looking for further research on the application of the AutoPhrase model.

References

[1] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss and Jiawei Han, *Automated Phrase Mining from Massive Text Corpora*, TKDE, doi:1702.04457, 2017

[2] Elliott Wave Principle: Key to Market Behavior by A.J. Frost Robert R. Prechter, Jr. Published by New Classics Library. ISBN 978-0-932750-75-4

[3] Joseph E. Granville, *Granville's New Strategy of Daily Stock Market Timing for Maximum Profit*, Prentice-Hall, Inc., 1976. ISBN 0-13-363432-9

[4] Carl B. Mcgowan, Junaina Muhammad, *The Relationship Between Price And Volume For The Russian Trading System*, International Business Economics Research Journal 2012, Volume 11, Number 9.

[5] Monica Lam, *Neural network techniques for financial performance prediction: integrating fundamental and technical analysis*, Decision Support Systems, Volume 37, Issue 4, September 2004, Pages 567-581.

[6] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky1, *On the Importance of Text Analysis for Stock Price Prediction.*, Language Resources and Evaluation Conference (LREC), 2014

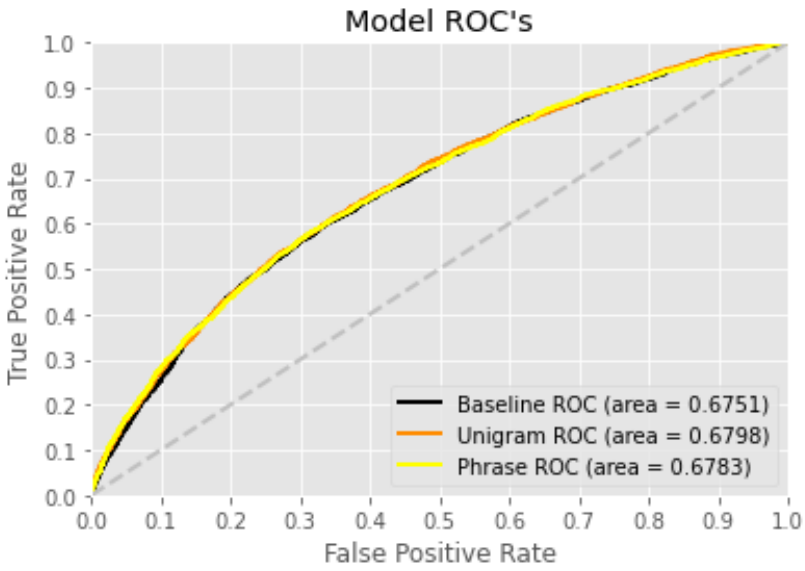
Model ROC's
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Baseline ROC (area = 0.6751)
Unigram ROC (area = 0.6798)
Phrase ROC (area = 0.6783)

Figure 4: ROC

	feature	importance
0	Surprise(%)	0.227200
1	price_change_365	0.048843
2	prev_vix_values	0.045149
3	price_change_7	0.035074
4	price_change_90	0.031560
5	price_change_30	0.029246
6	semiconductor	0.003361
7	results of operations and financial condition	0.003084
8	federal	0.003028
9	weak	0.002605
10	november	0.002462
11	world	0.001998
12	raising	0.001949
13	recovery	0.001944
14	ownership	0.001895
15	ubsi	0.001851
16	deregulation	0.001807
17	agilent	0.001791
18	nol	0.001782
19	bar	0.001754

Figure 5: Unigram Model Importance

	feature	importance
0	Surprise(%)	0.283263
1	price_change_365	0.063257
2	prev_vix_values	0.058757
3	price_change_7	0.046266
4	price_change_90	0.040840
5	price_change_30	0.038773
6	results of operations and financial condition	0.004311
7	energy services	0.003334
8	operating margin	0.002923
9	revenue growth	0.002915
10	fourth quarter	0.002722
11	risk factor	0.002713
12	annual report	0.002517
13	period end	0.002324
14	intellectual property	0.002311
15	financial statements and exhibits	0.002268
16	cash flow	0.002244
17	tax rate	0.002209
18	gaap measures	0.002158
19	production volumes	0.002090

Figure 6: Phrase Model Importance

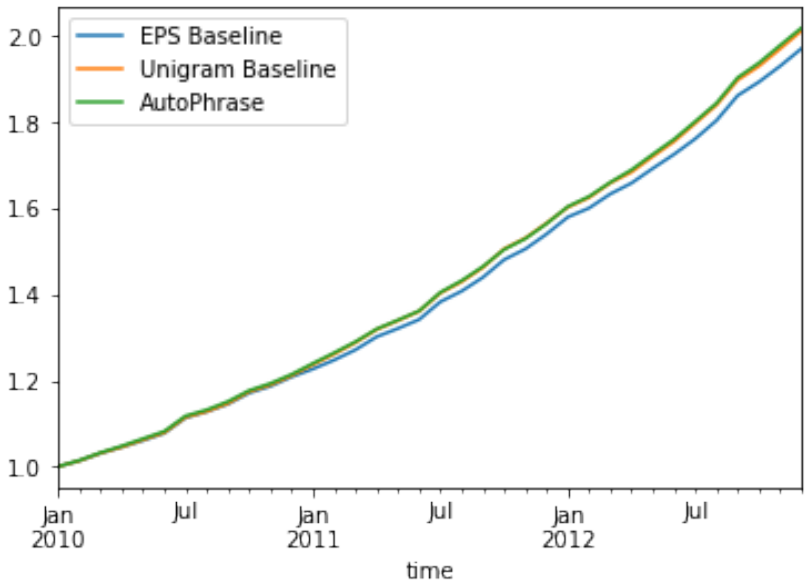

Figure 7: Rate of Return Curve