AnalyticsQueenBee Group Present

Follow the project on Github:
https://github.com/danielle707/

Stock Prediction and Trading Strategies using Social Media Sentiment Analysis

Group: AnalyticsQueenBee

0. Abstract

This project aims to investigate the prediction power of public sentiment on FANG (Facebook, Amazon, Netflix, Google) extracted from Reddit and New York times by comparing the prediction accuracy of different machine learning methods. We conclude that some simpler models, such as linear regression and random forest, which yield best prediction ability (63.16%) on stock return's growth trend with sentiment collected from Reddit, largely due to the high correlation between sentiment effects and stock price.

We hypothesize mean reversion for every stock. In analysis, we employ comprehensive machine learning approaches with multiple sentiment source including New York Times, Reddit and VIX index. Our results show that public data source presents polarized sentiments towards negative emotions such as fear and anger (relative importance 0.12). And Our results achieve improvement in every measured metric, especially in rolling-window trading strategies which achieve at most 70% monthly return on Amazon stock using linear model, and average return 20% for all four stocks using Reddit data source.

1. Data Collecting-Web Scraping

1.1 Reddit

We choose Reddit to represent public opinion. This clearly differs from classical official news websites like New York Times, where editors choose which content appears.

timestamp	upvote_ratio	score	comms_num	anticipation	sadness	joy	negative	trust	positive	surprise	disgust	anger	fear
2018-11-27	0.700000	7.000000	9.000000	0.006211	0.006211	0.000000	0.044275	0.006211	0.031454	0.000000	0.031852	0.000000	0.119197
2018-11-28	0.650000	38.000000	11.833333	0.096002	0.010329	0.001753	0.066249	0.031820	0.045100	0.002435	0.043672	0.006209	0.066567

As we need company related news, we chose the subreddit for our four companies (i.e. Amazon, Google, Netflix, and Facebook) as our main source of public opinions. To retrieve the required data, we use reddit API and package praw in Python. The data includes submission title, article and comments, a link to an online resource, the date the submission was published on, the absolute score of votes and the number of comments. Next, we write a second crawler which is able to download the news texts from the linked websites. We end up with 5404 articles from Nov 26th 2018 and Nov 11th 2019 and obtained several different news for each trading day.

1.2 New York Times

We choose New York Times as a contrast news source compared to Reddit. We chose them as a baseline, since the New York Times is considered as an important source for independent business-related news. To obtain the news from the New York Times website we have written another crawler using API, that iterates over the news archive of the New York Times and downloads all news data, including the news text and the date of publication. Using this method, we were able to collect 404 New York Times articles published between Nov 26th 2018 and Nov 11th 2019.

week	anticipation	sadness	joy	negative	trust	positive	surprise	disgust	anger	fear	modified_positive	modified_negative
2018-45	0.016256	0.002347	0.000000	0.004762	0.004695	0.022546	0.014351	0.009523	0.011937	0.007176	0.043497	0.033398
2018-46	0.017212	0.007583	0.004135	0.009544	0.005409	0.026501	0.006926	0.000000	0.016657	0.001493	0.053256	0.027693

1.3 Stock Price

We use Standard and Poor's 500 (S&P 500) and the four companies historical stock price from Nov 26th 2018 and Nov 11th 2019 obtained from Yahoo! Finance. We calculated the

log-return cased on adj. closing price for every trading day. The S&P 500 index is considered as

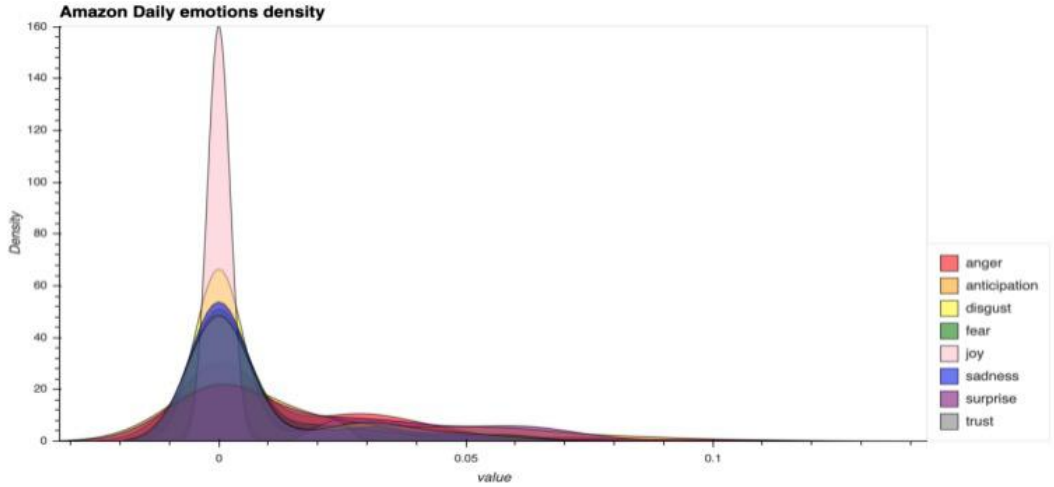
Date	Open	High	Low	Close	Adj Close	Volume	log_rt	direction	day	weekinfo	week
2018-11-26	1539.000000	1584.810059	1524.219971	1581.329956	1581.329956	6257700	NaN	1	2018-11-26	(2018, 48, 1)	2018-48
2018-11-27	1575.989990	1597.650024	1558.010010	1581.420044	1581.420044	5783200	0.000057	1	2018-11-27	(2018, 48, 2)	2018-48

one of the most important financial indicators worldwide. We chose this index since 54% of Reddit's visitors are come from the United States according to a user survey on Reddit.

2. Data Visualization and Variable Explanation

Wordclouds of all companies revolves around their geographical location and their major service, shed little insight on people's opinions. We conduct Exploratory Data Analysis (EDA) on both Reddit and New York Times sentiment data source. From density plot of different emotions, we find joy, most of the time, is absent from public news. Whereas negative emotions such as fear and anger present larger portions of emotions. The mean values of aggregated negative and positive sentiments are another evidence of this polarized distribution. All four companies demonstrate similar trend in public emotions, largely due to the industry and economic cycle.

To further explore the trend of public sentiment and stock growth, we plot all the emotions and stock price all together. The bokeh scatterplot shows no single sentiment with similar trends, indicating a more complex relationship or combine sentiment effect. In the following context, we aim to discover that relationship.



3. Baseline-Fundamental Analysis

3.1 GARCH

GARCH is a statistical modeling technique used to help predict the volatility of returns on financial assets. We first draw the volatility of each company. We can see that volatility become more volatile during financial and less volatile during steady economic growth. After conducting regression on the stock price and three attributes, the accuracy for Google, Facebook, Amazon and Netflix are 44%, 51%, 49%,50% respectively. (in appendix 5)

3.2 Fama-French Three Factor Model

The Fama French 3-factor model is an asset pricing model that expands on the capital asset pricing model by adding size risk and value risk factors to the market risk factors. After conducting regression on the stock price and three attributes, the accuracy for Google, Facebook,

Amazon and Netflix are 74%, 59%, 73%,94% respectively. Although the the Fama-French three factor model yield accuracy of 65%, its predictive powers still calls into question. According to the efficient market hypothesis (EMH), it is impossible to use publicly available information to yield higher profits, as information become part of the market in the very moment they are shared. Therefore, with the public financial information being well-explored, the predictive power of fundamental model may dwarf in front of the sentimental analysis model. (in appendix 5)

4. Machine Learning and Sentimental Analysis

In this analysis we evaluate different machine learning algorithms to predict stock market movements using financial news and comments posted on New York Times and Reddit. Specifically, we analyze all submissions in the subreddit corresponding to each company and articles in New York Times that have been posted between Nov 26th 2018 and Nov 11th 2019. We determine the sentiment variables in these news articles, and then, investigate the predictive power of them.

4.1 Sentimental Analysis

As for the New York Times, we organized the data into article title and article abstract; as for Reddit, we organized the data into upvote, score, discussion number, and discussion content. We then conduct sentiment analysis using NRC on New York Times and Reddit data, converting labeled articles into 2 sentiment and 8 emotions. Then we construct two sets of sentiment variables. The first one consists of all sentiments and emotions we get from NRC; and in the second set, we convert all sentiments into Modified Positive (Positive + joy + trust + anticipation) and Modified Negative (Negative + disgust + anger + fear) to serve as our explanatory variables. All these two sets are also containing the upvote, score, and discussion number when we analysis the data with respect to the Reddit data.

4.2 Machine Learning Models

After analyzing the sentiments in the collected news articles, our target variable is a classification variable, given by the direction of the stock market on the corresponding trading day (i.e. we denote positive log-return as 1 and negative log-return as 0). We define the threshold as the prediction mean. We then apply the following machine learning methods:

1. linear regression model

   Since the prediction results in linear regression model will not be in binary form, we classify it in to {0,1} by the threshold which is equal to the mean of all prediction results;
2. Decision tree: Classification trees

   A decision tree is decision support tool that uses a tree-like model of decisions and their possible consequences, which is a wonderful fit for our problem since it classified the prediction results.
3. Random forest

   We also use it to show the most important variable (i.e. sentiment) affecting the stock prices.
4. SVM: (1) linear SVM; (2) non-linear SVM: sigmoid
5. Neural Network: (1) linear NN: 1 layer; (2) non-linear NN: we test 2-4 layers, and find 3 layers. are the best in our data set.
6. Logistic Regression

4.3 Model Evaluating

We choose Amazon as our testing company, and test mainly 4 kinds of explanatory variables

sets, with modified emotions and all sentiments corresponding to Reddit and New York Times. Then, we construct a score form which contains the precision, true negative rate, f score, and accuracy in order to evaluate our models. We also plot the ROC curve for each model to better visualize our evaluation. Furthermore, we evaluate the importance of all sentiments and modified emotions and find that fear and modified negative always be the leading factors to the stock prices. (in Appendix 6-9, score form, ROC curve, and features importance)

Following these descriptions, we achieve an accuracy of 60% in predicting the future direction of the stock market. And based on our evaluation factors, we notice that linear regression and SVM (sigmoid) always be the most accuracy two models no matter we use all sentiments or use modified emotions for either Reddit or New York Times data. This may because we have only 5 to 10 explanatory variables which is smaller to fit some complex models like Neural Networks and Random Forest models with several layers. In addition to this, we notice that using all emotions generate a higher accuracy in both accuracy and precision than using modified emotions. (in Appendix 6-9, score form)

Moreover, we find our predictive accuracy is higher when we use Reddit data comparing to analyzing financial news (NYtimes) alone. This tells us using public information for predicting stock price is more robust than using some official news data (in Appendix 6-9)

5. Trading Strategy

5.1 Strategy Construction

Based on our model evaluation above, we choose to use linear regression model, random forest, SVM with sigmoid and logistic models to construct our trading strategies. Since New York times data is weekly basis and do not have enough number of data, we just use Reddit data since it is daily basis. All sentiments and modified emotions are both test in our strategies.

We then build our trading strategy based on rolling window methods. We choose a rolling window size m equal to the 70% of the whole data set and use each window to fit the machine learning models we chose to predict the stock price moving direction on the day after the training set. The first rolling window contains observations for period 1 through m, the second rolling window contains observations for period 2 through $m + 1$, and so on. According to our prediction, if price goes up, buy at previous day; if price goes down, sell at previous day.

5.2 Practical Trading Result

We apply the trading strategies to 4 companies (Amazon, Google, Netflix, and Facebook). From Appendix 10-13, we plot the cumulative returns for each prediction models and the baseline which is the raw returns, such that we do not do any but-sell actions for the stock. We also construct the evaluation form which contains the annualized-return, volatility and profit rate.

Our strategies based on linear regression, random forest and SVM models all generate higher return than raw data, while logistic based strategies are poorly performed in some cases. In addition, strategy using all eight sentiments are more reliable than the ones using modified emotions. Considering each firm, Amazon and Google generate higher and more stable returns may because they have higher discussion rate on Reddit and they have earlier time to market, especially for Amazon. The form below shows the annualized-return for each trading strategy on the four company. Further plots and forms are in Appendix 10-13.

	Amazon		Google		Netflix		Facebook	
	all emotions	modified emotions	all emotions	modified emotions	all emotions	modified emotions	all emotions	modified emotions
raw data	-37.13		12.15		-11.67		4.89	
linear model	146.64	232.38	82.63	28.69	42.74	21.15	8.53	3.76
random forest	276.35	215.33	8.37	72.53	16.07	21.81	5.41	7.27
SVM_sigmoid	57.44	46.15	89.06	29.77	18.25	18.43	6.03	6.34
logistic	172.61	190.07	-20.41	-21.94	6.14	1.07	-1.75	-4.54