Twitter 上的情绪如何预测股价走势

正文

贪婪和恐惧是股市的两大驱动力。事实证明，社交媒体信息中的积极和消极情绪，比如Twitter，可用于预测股票价格的日常变动或走势。

尽管新闻肯定会影响股市价格，但公众情绪状态也可能发挥同样重要的作用。**我们从心理学研究中得知，情感和信息一样，在人类的决策过程中扮演着重要的角色。行为金融学进一步证明，金融决策在很大程度上是由情绪驱动的。**因此我们有理由假设，公众情绪能够像新闻一样推动股市的价格。

这里有一些研究可供大家参考：

论文地址：https://arxiv.org/pdf/1010.3003.pdf

Twitter mood predicts the stock market.

Johan Bollen[1,*], Huina Mao[1,*], Xiao-Jun Zeng[2].

*: authors made equal contributions.

***Abstract*—Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.**

***Index Terms*—stock market prediction — twitter — mood analysis.**

I. Introduction

STOCK market prediction has attracted much attention from academia as well as business. But can the stock market really be predicted? Early research on stock market prediction [1], [2], [3] was based on random walk theory and the Efficient Market Hypothesis (EMH) [4]. According to the EMH stock market prices are largely driven by *new* information, i.e. news, rather than present and past prices. Since news is unpredictable, stock market prices will follow a random walk pattern and cannot be predicted with more than 50 percent accuracy [5].

There are two problems with EMH. First, numerous studies show that stock market prices do not follow a random walk and can indeed to some degree be predicted [5], [6], [7], [8] thereby calling into question EMH's basic assumptions. Second, recent research suggests that news may be unpredictable but that very early indicators can be extracted from online social media (blogs, Twitter feeds, etc) to predict changes in various economic and commercial indicators. This may conceivably also be the case for the stock market. For example, [11] shows how online chat activity predicts book sales. [12] uses assessments of blog sentiment to predict movie sales. [15] predict future product sales using a Probabilistic Latent Semantic Analysis (PLSA) model to extract indicators of sentiment from blogs. In addition, Google search queries have been shown to provide early indicators of disease infection rates and consumer spending [14]. [9] investigates the relations between breaking financial news and stock price changes. Most recently [13] provide a ground-breaking demonstration of how public sentiment related to movies, as expressed on Twitter, can actually predict box office receipts.

Although news most certainly influences stock market prices, public mood states or sentiment may play an equally important role. We know from psychological research that emotions, in addition to information, play an significant role in human decision-making [16], [18], [39]. Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood [19]. It is therefore reasonable to assume that the public mood and sentiment can drive stock market values as much as news. This is supported by recent research by [10] who extract an indicator of public anxiety from LiveJournal posts and investigate whether its variations can predict S&P500 values.

However, if it is our goal to study how public mood influences the stock markets, we need reliable, scalable and early assessments of the public mood at a time-scale and resolution appropriate for practical stock market prediction. Large surveys of public mood over representative samples of the population are generally expensive and time-consuming to conduct, cf. Gallup's opinion polls and various consumer and well-being indices. Some have therefore proposed indirect assessment of public mood or sentiment from the results of soccer games [20] and from weather conditions [21]. The accuracy of these methods is however limited by the low degree to which the chosen indicators are expected to be correlated with public mood.

Over the past 5 years significant progress has been made in sentiment tracking techniques that extract indicators of public mood directly from social media content such as blog content [10], [12], [15], [17] and in particular large-scale Twitter feeds [22]. Although each so-called *tweet*, i.e. an individual user post, is limited to only 140 characters, the aggregate of millions of tweets submitted to Twitter at any given time may provide an accurate representation of public mood and sentiment. This has led to the development of real-time sentiment-tracking indicators such as [17] and "Pulse of Nation"[1].

In this paper we investigate whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the stock market. We use two tools to measure variations in the public mood from tweets submitted

[1] http://www.ccs.neu.edu/home/amislove/twittermood/

论文地址：

https://link.springer.com/article/10.1057/s41265-016-0034-2

Journal of Information Technology
March 2018, Volume 33, Issue 1, pp 50–69 | Cite as

More than just noise? Examining the information content of stock microblogs on financial markets

Authors | Authors and affiliations

Ting Li, Jan van Dalen, Pieter Jan van Rees

Research Article
First Online: 07 March 2017

论文地址：http://blueanalysis.com/iulianserban/Files/twitter_report.pdf

Prediction of changes in the stock market using twitter and sentiment analysis

Iulian Vlad Serban, David Sierra González, and Xuyang Wu
University College London

***Abstract*— Twitter is an online social networking and microblogging service with over 200m monthly active users. Given this massive user base researchers have tried to mine the derived vast source of data for different purposes. In this work, we investigate the relationship between the market indicators for three companies (IBM, Intel and General Electric) and the volume of tweets mentioning their names or stock symbols. We consider additionally other factors, such as the predicted sentiment of the tweets, the number of followers/friends of the users and the presence of links on the tweets. With all this information a predictor is trained for each company to estimate the changes in the stock market price. An exhaustive feature selection procedure was performed, showing that the most correlated features with the stock market indicators were the number of tweets weighted by the number of friends. After selecting the four most correlated Twitter related features, and together with the stock market indicators at previous timesteps, six different approaches were studied as predictive models, namely, linear regression considering only the tweet counts, linear regression including sentiment features, non-linear regression considering higher-order interactions between the sentiment-based features and the stock market indicators, and the LASSO regularized versions of the three models. All models performed consistently better than two benchmark models (constant and random prediction) for the three stocks, according to the mean absolute error and mean squared error metrics. This confirms the existance of predictive power in the Twitter features. However, no significative difference was observed between the models using sentiment features and those considering only the tweet counts.**

I. Introduction

Twitter is a free microblogging service founded in 2006 by Jack Dorsey and Biz Stone. It enables users to send and read *tweets*, which are text messages limited to 140 characters. Registered users of Twitter are able to read and post tweets via the web, SMS or mobile applications. The user base of Twitter surpassed the 200 million active users in December 2012 [1].

With such an impressive user base researchers have become interested in mining Twitter data to extract patterns and trends. Understanding how and why people tweet seems like a reasonable first step. Twitter is currently being used for *daily chatter*, *conversations*, *sharing information/URLs*, and *reporting news*; and its users can be classified into the groups such as *information sources*, *friends*, and *information seekers* [2].

Already in 2004 there was an study correlating web buzz and stock market [3]. In this work, Antweiler and Frank analyse how Internet stock message boards are related to stock markets. They conclude with the thought that there is financially relevant information present. In 2006, blog sentiment was used to predict movie sales [4].

More recent research suggests that online social media (blogs, Twitter feeds, etc.) can predict changes in various economic and commercial indicators [5]. In particular, the mood of the tweets, when classified into the mood dimensions *Calm*, *Alert*, *Sure*, *Vital*, *Kind* and *Happy*, have been shown to be significantly correlated with the Dow Jones Industrial Average index (DJIA). In [6], Bollen et al. show that sentiment analysis of Twitter posts over a period of 5 months is correlated with fluctuations in macro-social and -economic indicators in the same time period. A similar approach is taken in [7], where the positive and negative mood of tweets on Twitter is analysed and compared with stock market indices such as Dow Jones, S&P 500, and NASDAQ over a period of 5 months. They found that the number of positive tweets is much higher than that of negative ones, more than double on average. However, the mood indicators (both positive and negative) proved to be always negatively correlated with DJIA, NASDAQ and S&P500.

In 2012, Mao et al. investigated the correlations between the number of tweets that mention S&P 500 stocks and the stock indicators [8]. They applied a simple linear regression model with the tweet counts as exogenous input (independent variable) to predict the stock market indicators. Testing the model on a short period of 17 days, they reported an accuracy of 68% in predicting the direction of change in the daily closing price at stock market level.

In this work, we will address the following research questions:

- How should we analyse and interpret the sentiment of thousands of emotional tweets?
- What is the intrinsic relationship between emotional tweets and stock market?
- What class of models can we expect to perform well on such stock price and trading volume prediction across several stocks?
- What metrics are suitable for evaluating a social media based model for stock market prediction?

The project workflow we have established is shown in figure 1. Initially, the dataset is parsed and processed. As the next step, we analyse the sentiment of each English tweet related to the selected companies. We use the sentiments and related information from the tweets to extract a set of features,

论文地址：http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

Stock Prediction Using Twitter Sentiment Analysis

Anshul Mittal
Stanford University
anmittal@stanford.edu

Arpit Goel
Stanford University
argoel@stanford.edu

ABSTRACT

In this paper, we apply sentiment analysis and machine learning principles to find the correlation between "public sentiment" and "market sentiment". We use twitter data to predict public mood and use the predicted mood and previous days' DJIA values to predict the stock market movements. In order to test our results, we propose a new cross validation method for financial data and obtain 75.56% accuracy using Self Organizing Fuzzy Neural Networks (SOFNN) on the Twitter feeds and DJIA values from the period June 2009 to December 2009. We also implement a naive protfolio management strategy based on our predicted values. Our work is based on Bollen et al's famous paper which predicted the same with 87% accuracy.

1. INTRODUCTION

Stock market prediction has been an active area of research for a long time. The Efficient Market Hypothesis (EMH) states that stock market prices are largely driven by new information and follow a random walk pattern. Though this hypothesis is widely accepted by the research community as a central paradigm governing the markets in general, several people have attempted to extract patterns in the way stock markets behave and respond to external stimuli.

In this paper, we test a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals affect their decison making process, thus, leading to a direct correlation between "public sentiment" and "market sentiment". We perform sentiment analysis on publicly available Twitter data to find the public mood and the degree of membership into 4 classes - Calm, Happy, Alert and Kind (somewhat like fuzzy membership). We use these moods and previous days' Dow Jones Industrial Average (DJIA) values to predict future stock movements and then use the predicted values in our portfolio management strategy.

Related work Our work is based on Bollen et al's strategy [1] which received widespread media coverage recently. They also attempted to predict the behavior of the stock market by measuring the mood of people on Twitter. The authors considered the tweet data of all twitter users in 2008 and used the OpinionFinder and Google Profile of Mood States (GPOMS) algorithm to classify public sentiment into 6 categories, namely, Calm, Alert, Sure, Vital, Kind and Happy. They cross validated the resulting mood time series by comparing its ability to detect the public's response to the presidential elections and Thanksgiving day in 2008. They also used causality analysis to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. The authors used Self Organizing Fuzzy Neural Networks to predict DJIA values using previous values. Their results show a remarkable accuracy of nearly 87% in predicting the up and down changes in the closing values of Dow Jones Industrial Index (DJIA).

The rest of the paper is organized as follows. The second section briefly discusses our general approach towards solving the problem and the following sections discuss the individual components in greater detail. In Section 3, we briefly discuss the dataset that we have used for this paper and data preprocessing measures adopted. Section 4 discusses the sentiment analysis technique developed by us for the purpose of this paper. Section 5 includes in detail, the different machine learning techniques to predict DJIA values using our sentiment analysis results and presents our findings. In Section 6, we use the predicted values and devise a naive strategy to maintain a profitable portfolio.

2. ALGORITHM

The technique used in this paper builds directly on the one used by Bollen et al. [1]. The raw DJIA values are first fed into the preprocessor to obtain the processed values. At the same time, the tweets are fed to the sentiment analysis algorithm which outputs mood values for the four mood classes for each day. These moods and the processed DJIA values are then fed to our model learning framework which uses SOFNN to learn a model to predict future DJIA values using them. The learnt model as well as the previous DJIA and mood values are used by the portfolio management system which runs the model to predict the future value and uses the predicted values to make appropriate buy/sell decisions. Figure 1 shows a brief flow diagram of our technique.

The following sections discuss each component of our technique in greater detai

3. DATASET

In this project, we used two main datasets-

- Dow Jones Industrial Average (DJIA) values from June 2009 to December 2009. The data was obtained using Yahoo! Finance and includes the open, close, high and low values for a given day.

本文整个分析过程都是基于 **Python 编写。**

普及一个知识：

1、Twitter（推特）：是国外的一个社交网络及微博客服务的网站。

2、Tweet：是用户发到 Twitter 上的信息，为了接收或者发送 Tweets 首先要注册一个免费的 Twitter 帐号。

3、微博（MicroBlog）：是一个基于用户关系的信息分享、传播以及获取平台，用户可以通过 WEB、WAP 以及各种客户端组件个人社区，以 140 字左右的文字更新信息，并实现即时分享。

假设

今天的 Tweet 带有正面或负面情绪，并包含一个或几个 cashtags 可以影响股票明天的走势。如果今天负面情绪占主导地位，那么明天的股票价格预计会下跌，反之则会上涨。Twitter 账户的粉丝数量也是一个主要因素。一个账户的关注者越多，推文的影响力就越大，他们的情绪对股价的影响也越大。

cashtags 是什么?

Twitter 的一项功能允许用户点击股票代码，看看“Twitterverse”在说些什么，，比如$GOOG、$AAPL 或$FB。该系统的工作方式 Twitter 众所周知的#hashtags 相同。Cashtags 要求“$”后面跟着股票代码。

公众号补充：

一个通用标准$符号被纳入了 twitter 的官方标记（cashtag），Twitter 宣布这是包含了股票跟踪链接，用户点击股票信息便会显示到搜索页面上。

国内的雪球早已将$标记融入自己的微博服务中，且这些投资社区对$标记利用得更好。点击$标记后可显示出对应公司/股票的实时股价等交易信息及其他投资者对于这支股票的讨论。

数据集

从 2016 年 3 月 28 日到 2016 年 6 月 15 日，79 天内收集了大约 100 万条推文，其中提到了纳斯达克 100 指数成分股公司的 cashtags。这些数据由 followthehashtag.com 提供，这是一个 Twitter 搜索分析和商业智能工具。

https://www.followthehashtag.com/datasets/nasdaq-100-companies-free-twitter-dataset/

这里有两个带有 cashtags 的负面和正面推文的例子，分别代表苹果、谷歌和其他少数公司。

Date	Hour	Tweet content	Followers	Compound	neg	neu	pos
2016-06-15	09:44	#YouTube is built on the back of stolen content: Trent Reznor. $AAPL $GOOGL #MusicBiz	5172	-0.4939	0.176	0.824	0.0
2016-06-15	08:39	These 3 Stocks Will Surprise Investors This Earnings Season $ADBE $TTWO $ANET Also $AAPL $AMZN $FB $GOOGL	377	0.2732	0.0	0.884	0.116

在数据中的 100 只原始股票中，不得不因为各种数据特定的原因而减了 15 只，比如日期上的不一致，或者仅仅是因为关于 cashtags 的推文太少，也就是说，甚至连每天的推文都没有。排除在外的人包括 Apple，Tesla 和 Yahoo。

最终分析中包含推文最多的 cashtags 是（前 12 名）：

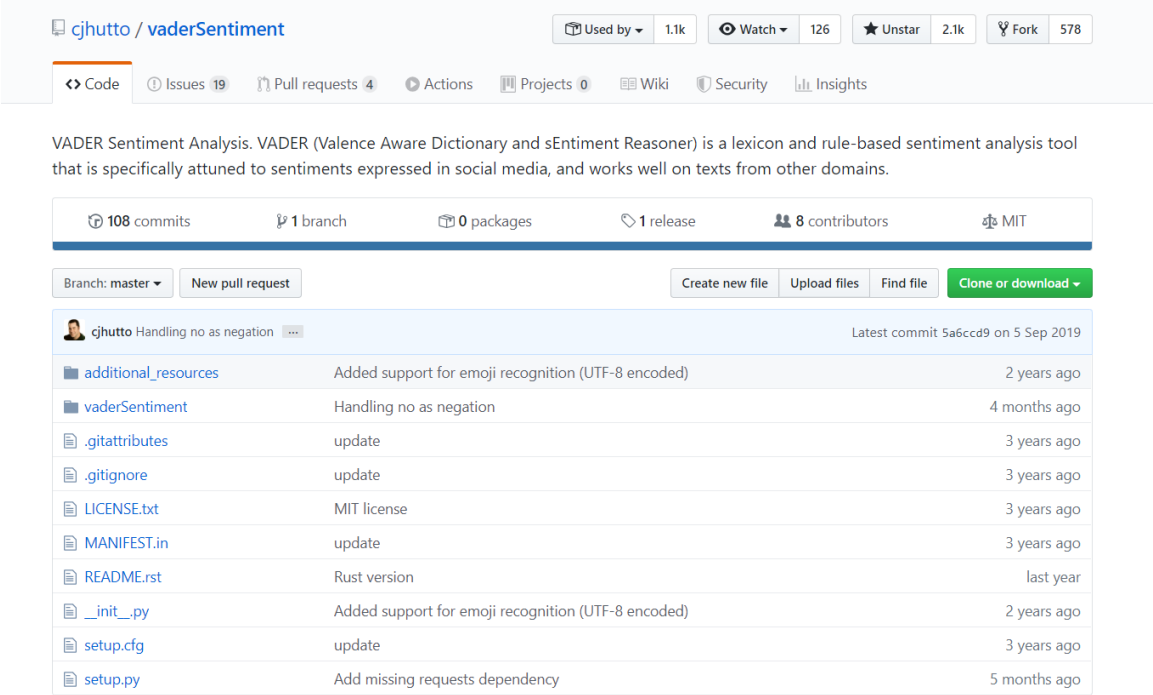
Company	Cashtag	Number of tweets
Facebook	$FB	93 898
Amazon	$AMZN	57 378
Microsoft	$MSFT	43 060
Alphabet Class A	$GOOG	37 642
Netflix	$NFLX	37 083
Alphabet Class C	$GOOGL	29 385
Gilead Sciences	$GILD	16 146
Starbucks	$SBUX	13 941
Cisco Systems	$CSCO	13 283
Nvidia	$NVDA	10 933

在这 79 天的时间里，100 只股票 cashtags 的平均推文数为 6446 条，即每只股票/cashtags 每天有 81 条推文。

衡量 tweets 上的情绪

为了提取每条 tweets 的情绪，我们使用了 **VADER**，这是一个现成的 Python 机器学习库，用于自然语言处理，特别适合阅读 tweets 的情绪。

地址：https://github.com/cjhutto/vaderSentiment



VADER 更注重大写字母的识别，还能识别俚语、感叹号和最常见的表情符号。情绪得分从极负（-1）到极正（+1），中性为 0。比如：

Text	Compound	Positive	Neutral	Negative
VADER is smart, handsome, and funny.	0.8316	0.746	0.254	0.0
VADER is VERY SMART, handsome, and FUNNY!!!	0.9342	0.767	0.233	0.0
VADER is not smart, handsome, nor funny.	-0.7424	0.0	0.354	0.646
Today SUX!	-0.5461	0.0	0.221	0.779
Today only kinda sux! But I'll get by, lol	0.5249	0.317	0.556	0.127

为 tweet 数据创建每日平均值

在将每条推文与其情绪相结合后，将其乘以该帐户的关注者数量。这样，在最终的模型中，更多“有影响力”账户的推文情绪将得到了更多的权重。在此之后，这些推文（平均每条 cashtags 有 6500 条）被压缩到 75 行，其中包括每条情绪的每日平均值，然后将其与相关股票的每日价格变化进行比较。

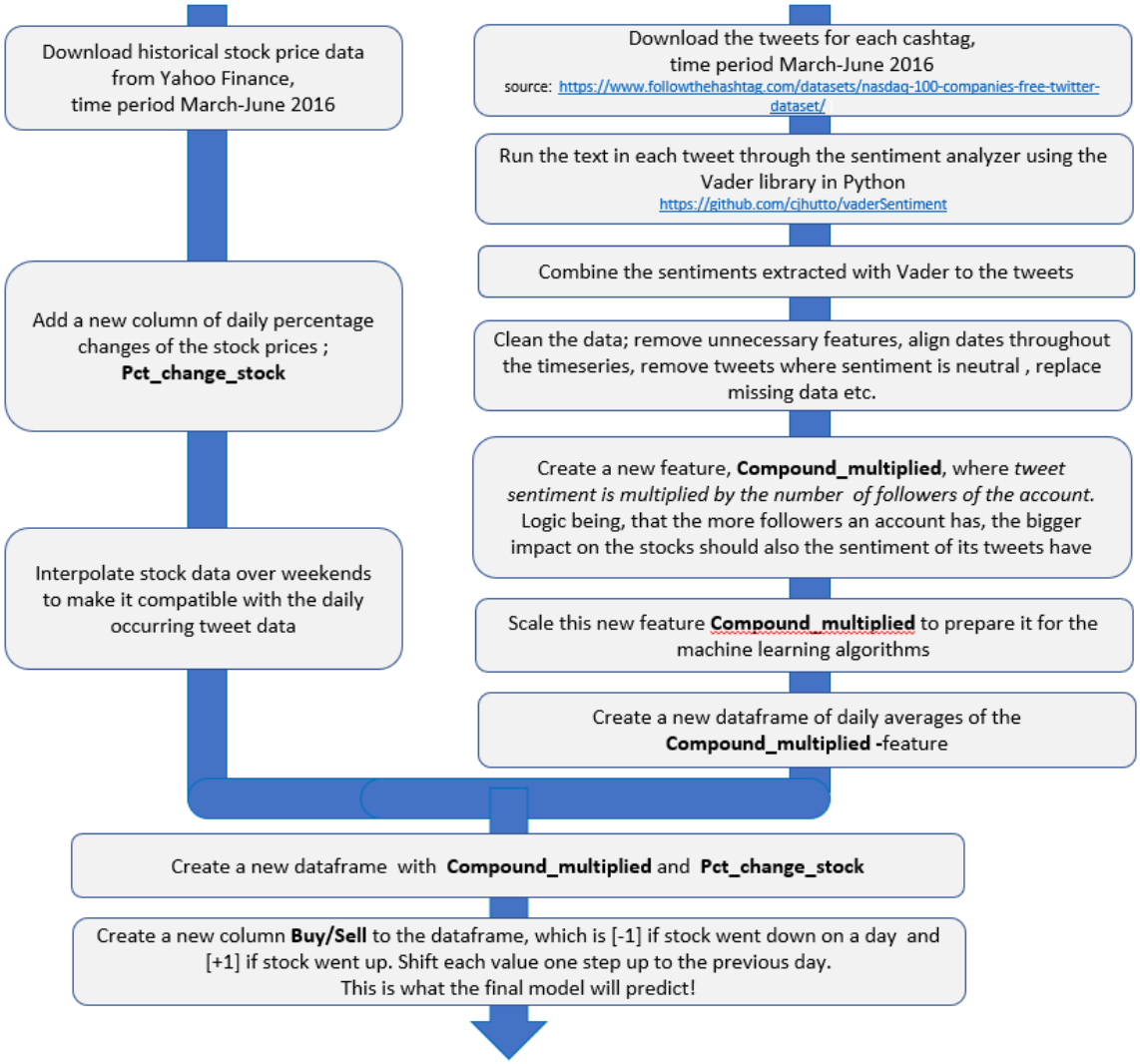
收集股票数据

使用 Python 的 pandas-datareader 库，从 Yahoo Finance 下载股票的每日数据。在股票数据中添加每日百分比变化列，并对周末缺失的数据进行插值之后，现在可以合并这两个数据集，即推文的情绪和股票的每日变化。

一个具有“Pct_change_stock”和“compound_multiply”两个特征的新 dataframe，以及一个添加标签数据列“Buy/Sell”，现在已经准备好在训练中使用。

第一部分流程图分析

股票数据（左箭头）Twitter 数据（右箭头）



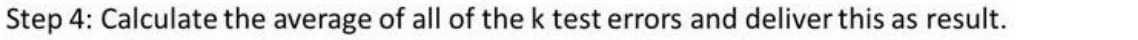
机器学习分类器

由于这是一个二元分类任务，即结果要么是“买入”，要么是“卖出”，因此我们使用了 6 种这样的算法：

- **KNN**
- **Logistic 回归**
- **支持向量机（SVM）**
- **朴素贝叶斯**
- **决策树**
- **随机森林**

训练/测试数据分割

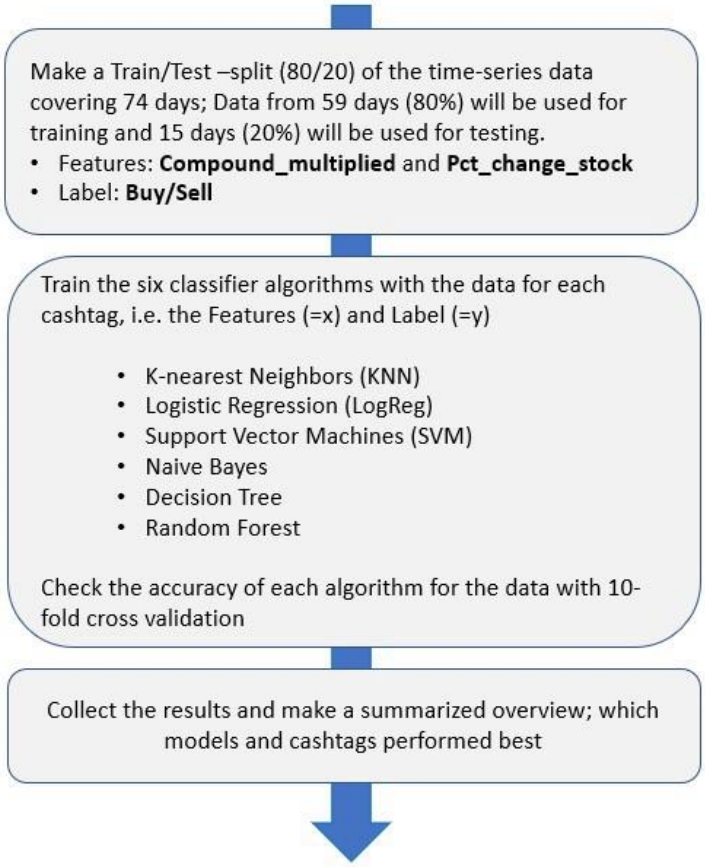
在 74 天可用的数据中，每只股票 59 天（80%）的数据用于训练，15 天（20%）的数据用于测试每种算法的准确性。



交叉验证

由于数据量有限，仅使用 20%的数据（15 天）和 80%的训练数据（59 天）进行测试可能不够有代表性。为了避免训练/测试分割不完全随机的可能性，对数据进行交叉验证，这样得到每个算法精度更具代表性的结果。训练数据进一步分成 10 个子集，每个子集都与其他 9 个子集进行测试。

第二部流程图分析



结果

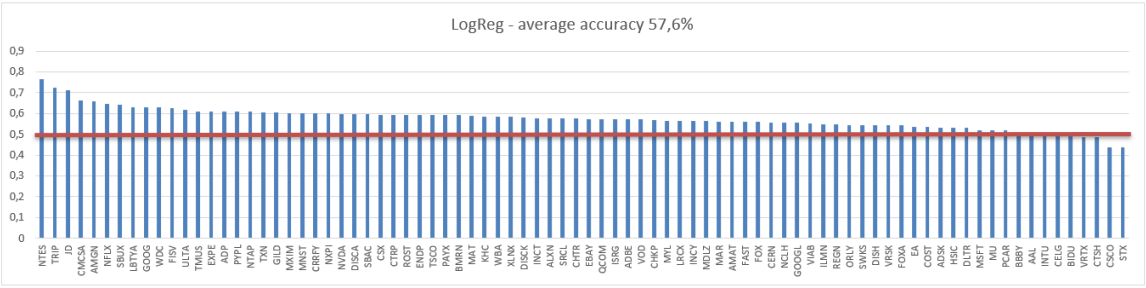
将 85 只股票分别通过 6 个二元分类器和 10 倍交叉验证后，结果如下。**平均每个分类器的准确率都在 50%以上。这意味着，推特上的情绪具有预测力，至少比抛硬币强。抛硬币的平均准确率为 50%，所以准确率超过 50%在一定程度上证明了模型获得“非凡”收益的能力。更重要的是，对于许多股票，模型的准确性/预测能力在 65-75%之间!**

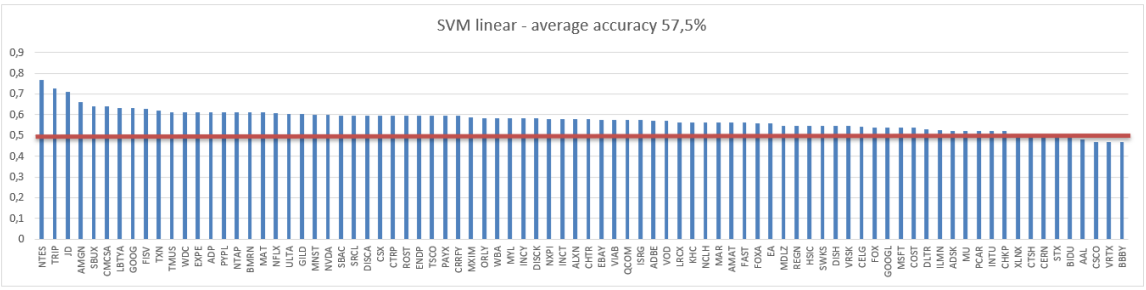
每个 cashtag 分类器的准确率（前 5 名）：

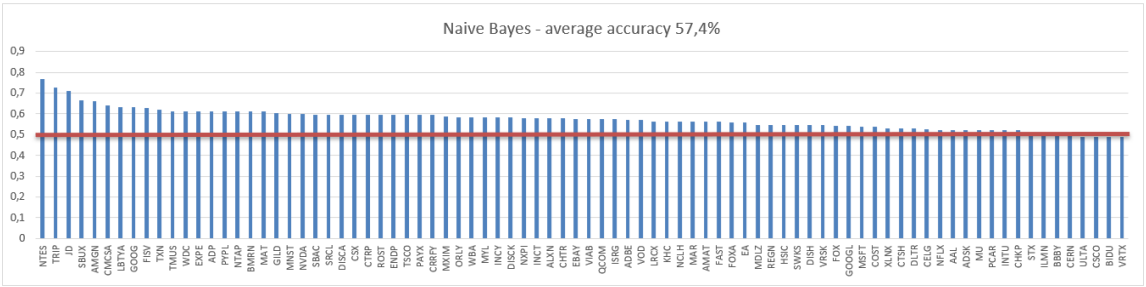
LogReg		SVM linear		Naive Bayes		KNN		Decision Tree		Random Forest	
Cashtag	accuracy	Cashtag	accuracy	Cashtag	accuracy	Cashtag	accuracy	Cashtag	accuracy	Cashtag	accuracy
NTES	76,6 %	NTES	76,6 %	NTES	76,6 %	FISV	70,7 %	MNST	72,7 %	FISV	71,3 %
TRIP	72,5 %	TRIP	72,5 %	TRIP	72,5 %	TMUS	68,3 %	SRCL	71,7 %	MNST	70,7 %
JD	71,1 %	JD	71,1 %	JD	71,1 %	CMCSA	66,5 %	FISV	71,0 %	SRCL	68,3 %
CMCSA	66,3 %	AMGN	66,0 %	SBUX	66,7 %	ILMN	65,8 %	REGN	67,9 %	GOOG	67,6 %
AMGN	66,0 %	SBUX	64,2 %	AMGN	66,0 %	DLTR	65,7 %	GOOG	67,6 %	SBUX	67,2 %

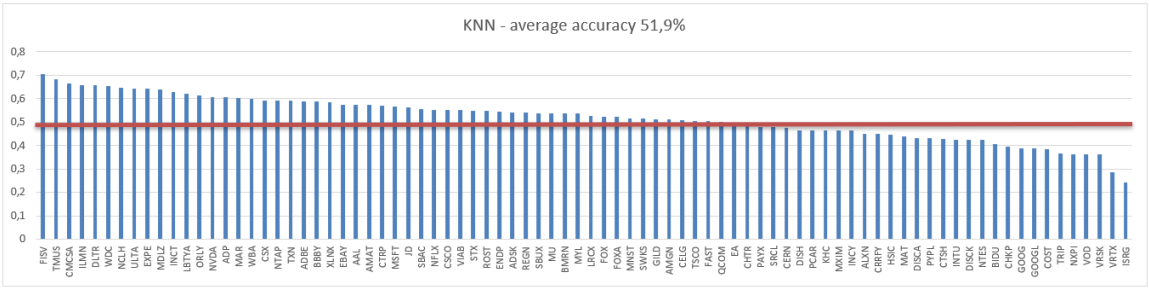
在下面的图表中，红线表示 50%的准确度限制。

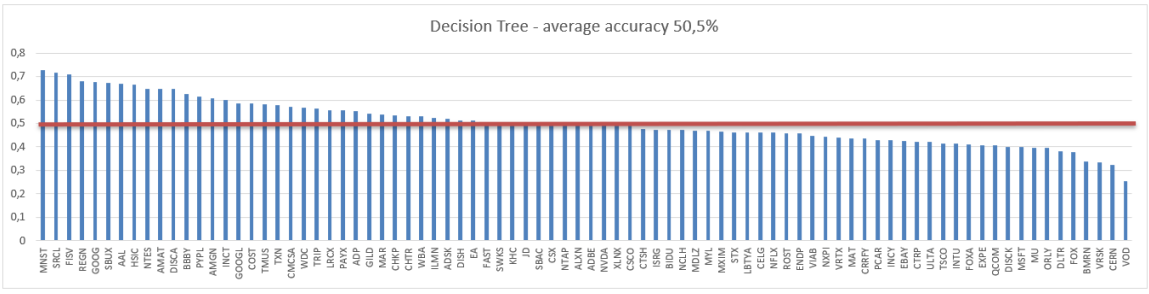
以下是所有分类器的平均准确率：

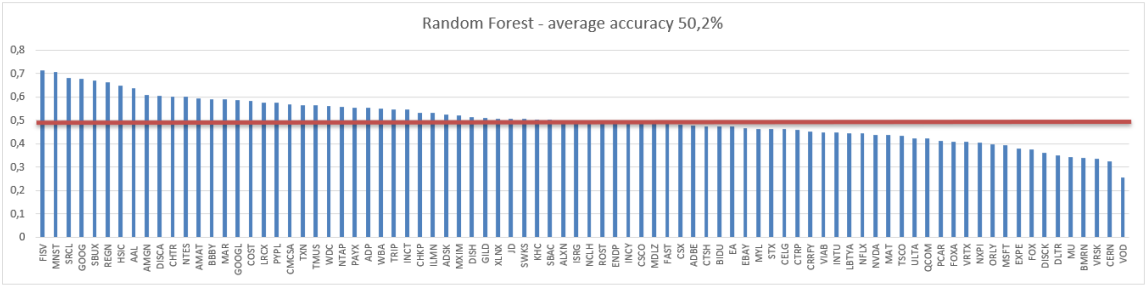












接下来，我们将简单买入持有策略的盈亏与使用模型实现的盈亏进行了比较。令我们惊讶的是，在为期四周的模拟交易中，大多数模型的利润都远超我们的预期！

下载 tweets

我们选择了纳斯达克的 8 只股票进行模拟，三月模拟交易的推文总数接近 7200，平均大约 800 每条股票的推文。

Company/stock	cashtag	Nr of tweets in March 2019	Average tweets per day
American Airlines Group	$AAL	1827	65
Automatic Data Processing Inc	$ADP	784	28
Cerner Corporation	$CERN	446	16
Expedia	$EXPE	519	19
Fiserv	$FISV	497	18
T-mobile US	$TMUS	846	30
Texas Instruments	$TXN	997	37
Western Digital	$WDC	1201	43

tweet 数据是通过使用其 Developer API“抓取”Twitter 而收集的。我们在 2016 年 3 月下载了所有包含 cashtags $AAL、$ADP、$CERN、$EXPE、$FISV、$TMUS、$TXN 和$WDC 的 tweets。

下载和准备其余的数据

前面我们详细解释了后续步骤的过程，下面简要的做一个回顾：

1、推文通过情绪分析算法运行，每个推文都有一个情绪；积极的，中性的或消极的。

2、每条推文都乘以该账户的关注者数量。这样，在最终的模型中，更“有影响力”账户的推文情绪就会得到更多的权重。

3、Tweet 数据被压缩到 28 行，包含每一个情绪的日平均，并与同期相关股票的日价格变化进行比较。

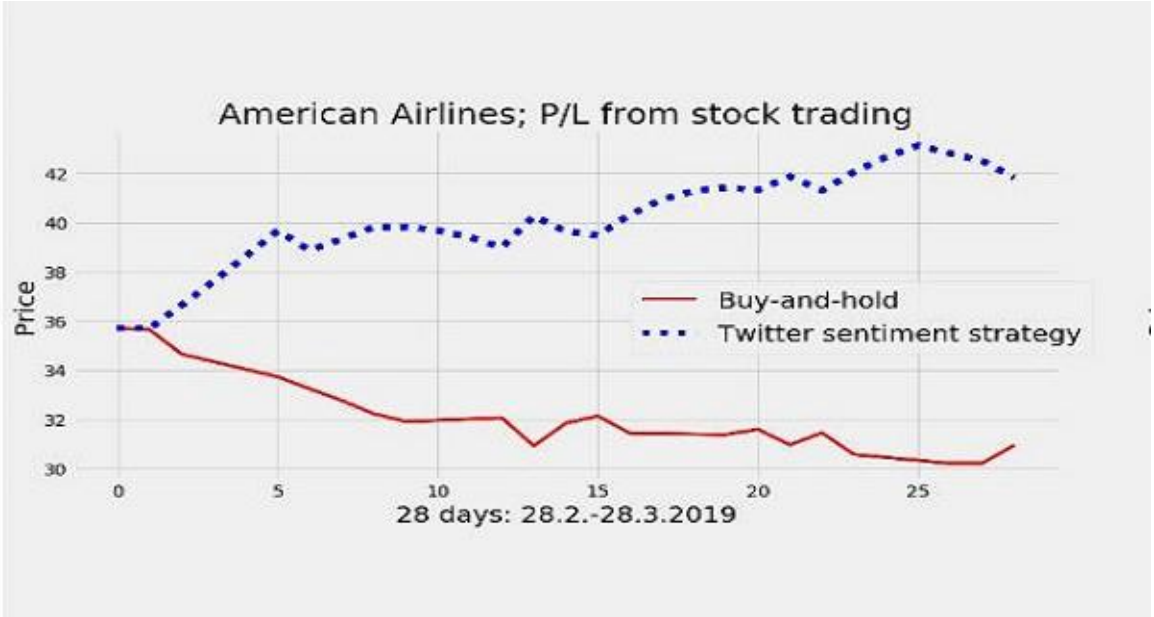
4、股票数据下载并添加“每日变化百分比”列中。

5、Tweet 和股票数据相结合，并添加一个标签列，即“买进或卖出”。这就是模型试图预测的内容。换句话说，基于今日推特情绪的预测值，预测一只股票应该在明天买进还是卖出?

然后通过比较买入持有策略与六种不同模型来使用这些数据集，每个每日预期的每日股票价格变动是使用模型预测的。

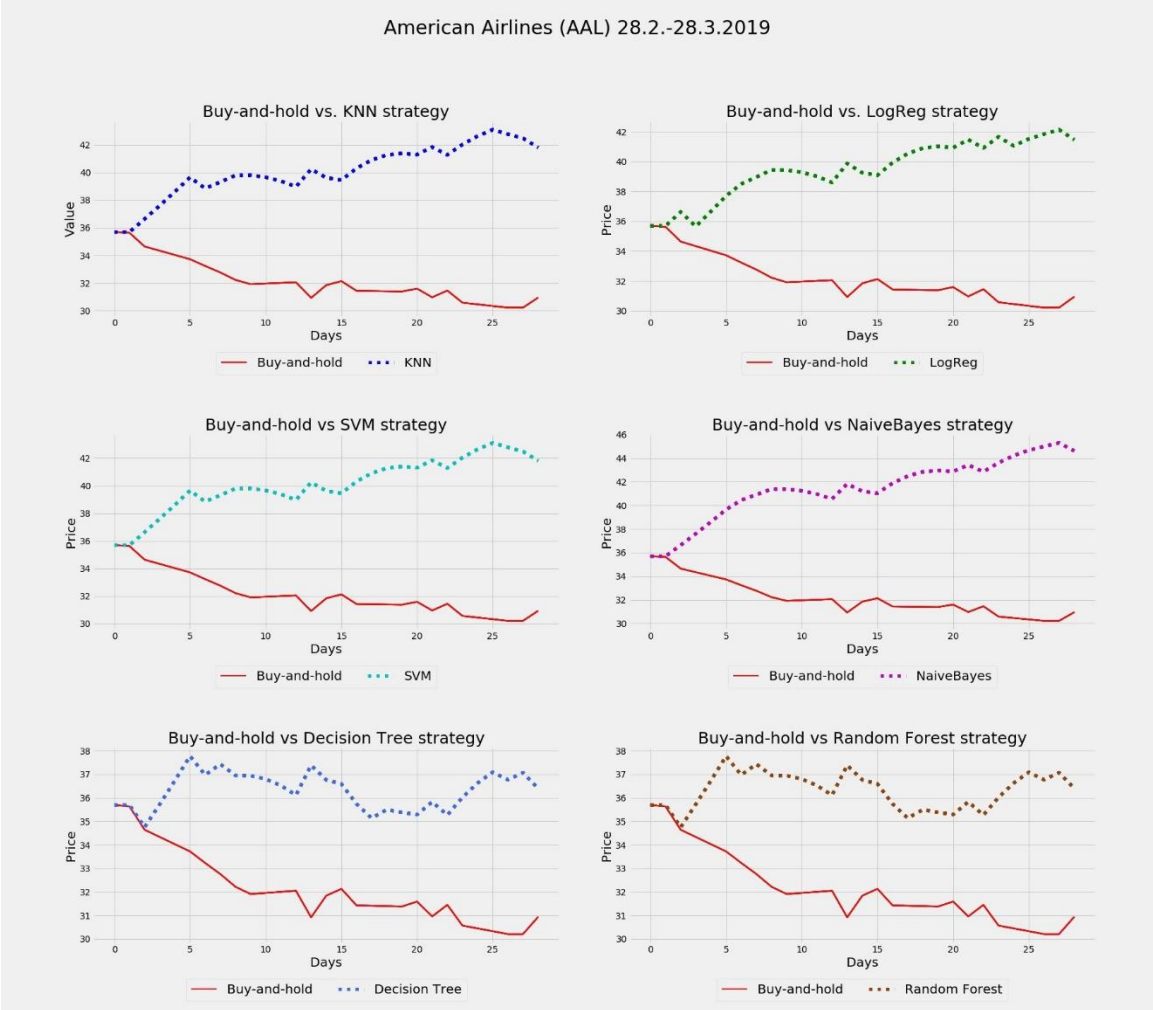
进行模拟交易 2019 年 3 月

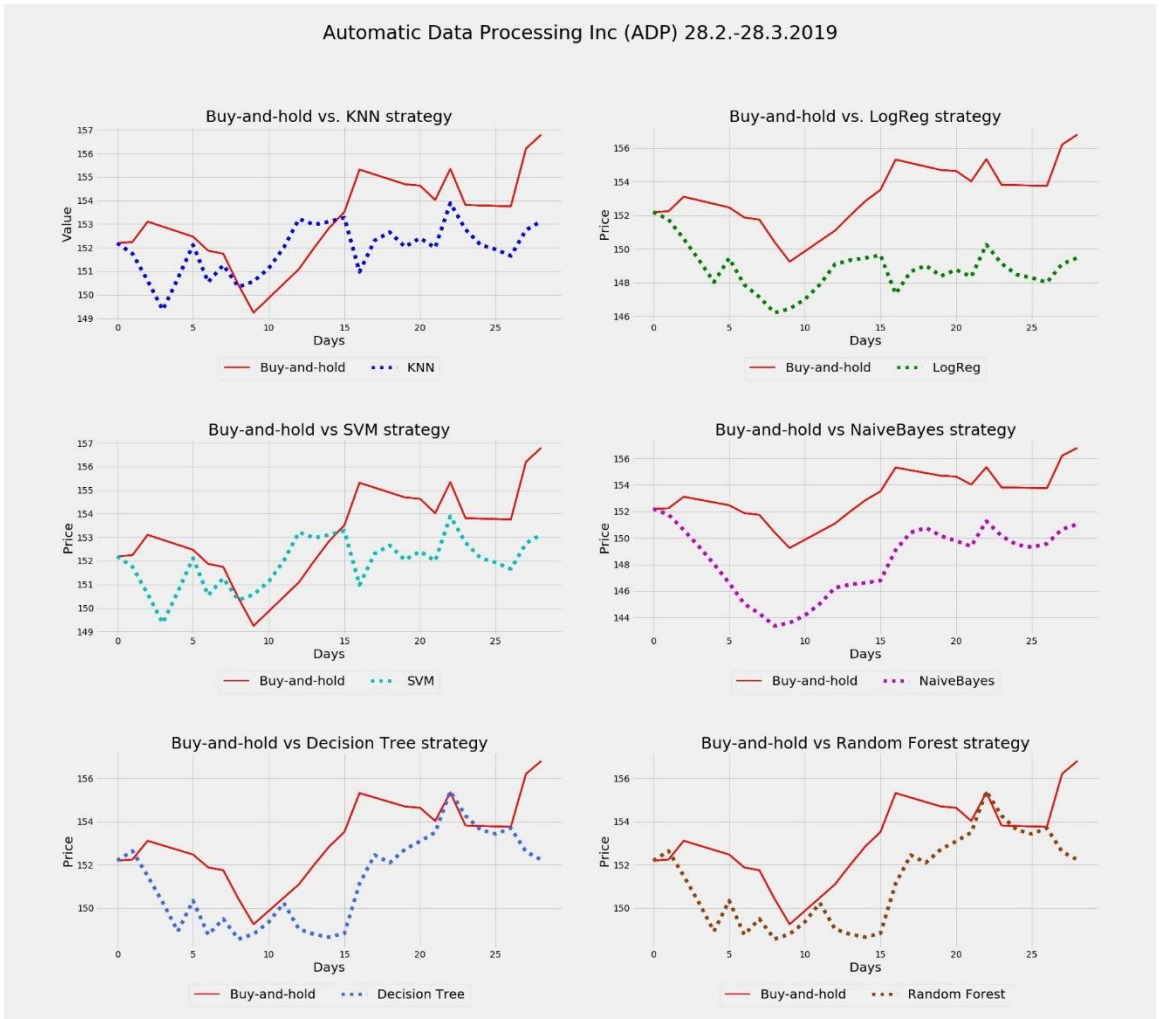
对 8 只股票分别采用买入并持有策略，与其他 6 种基于二分类算法策略进行比较。

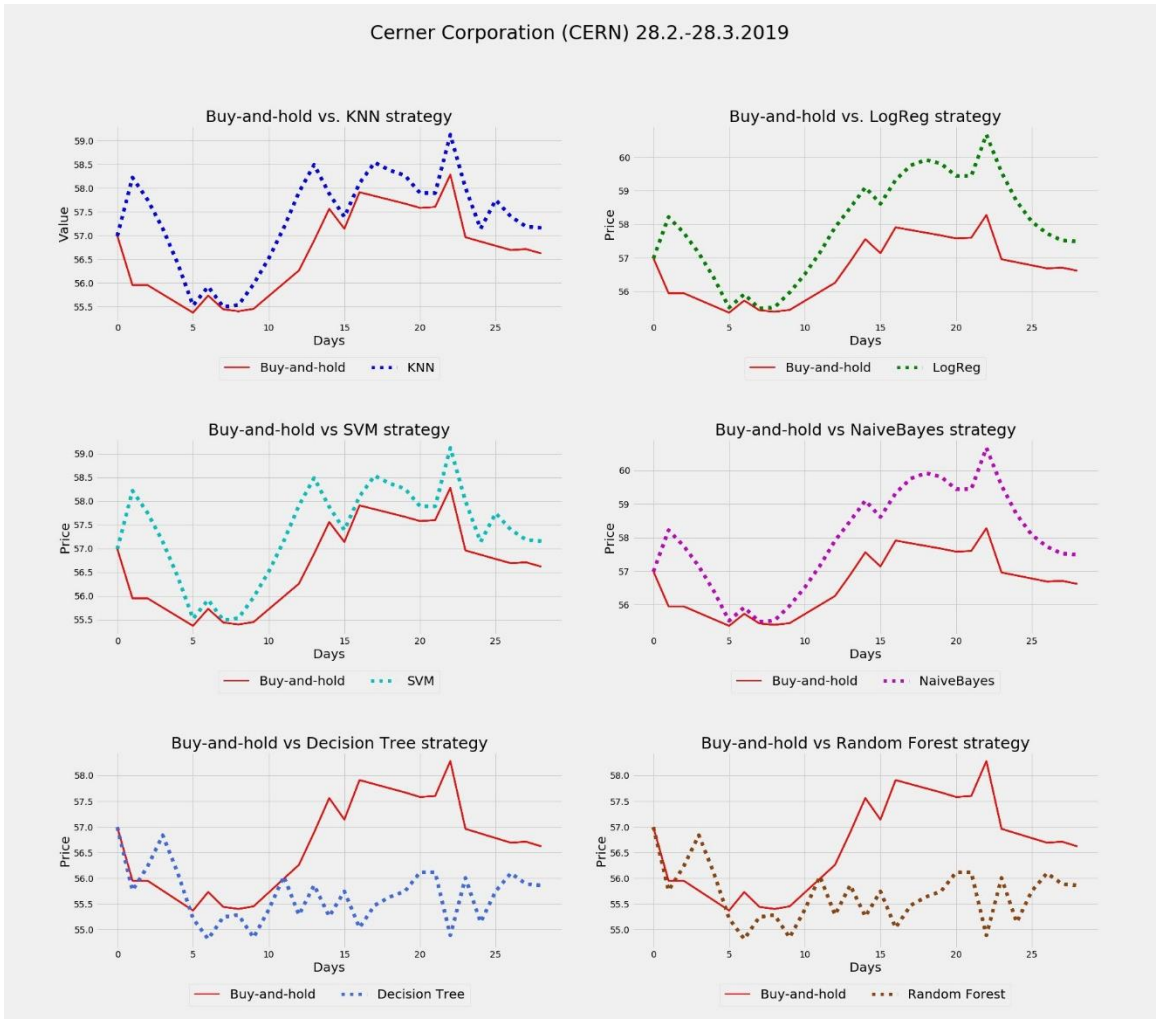


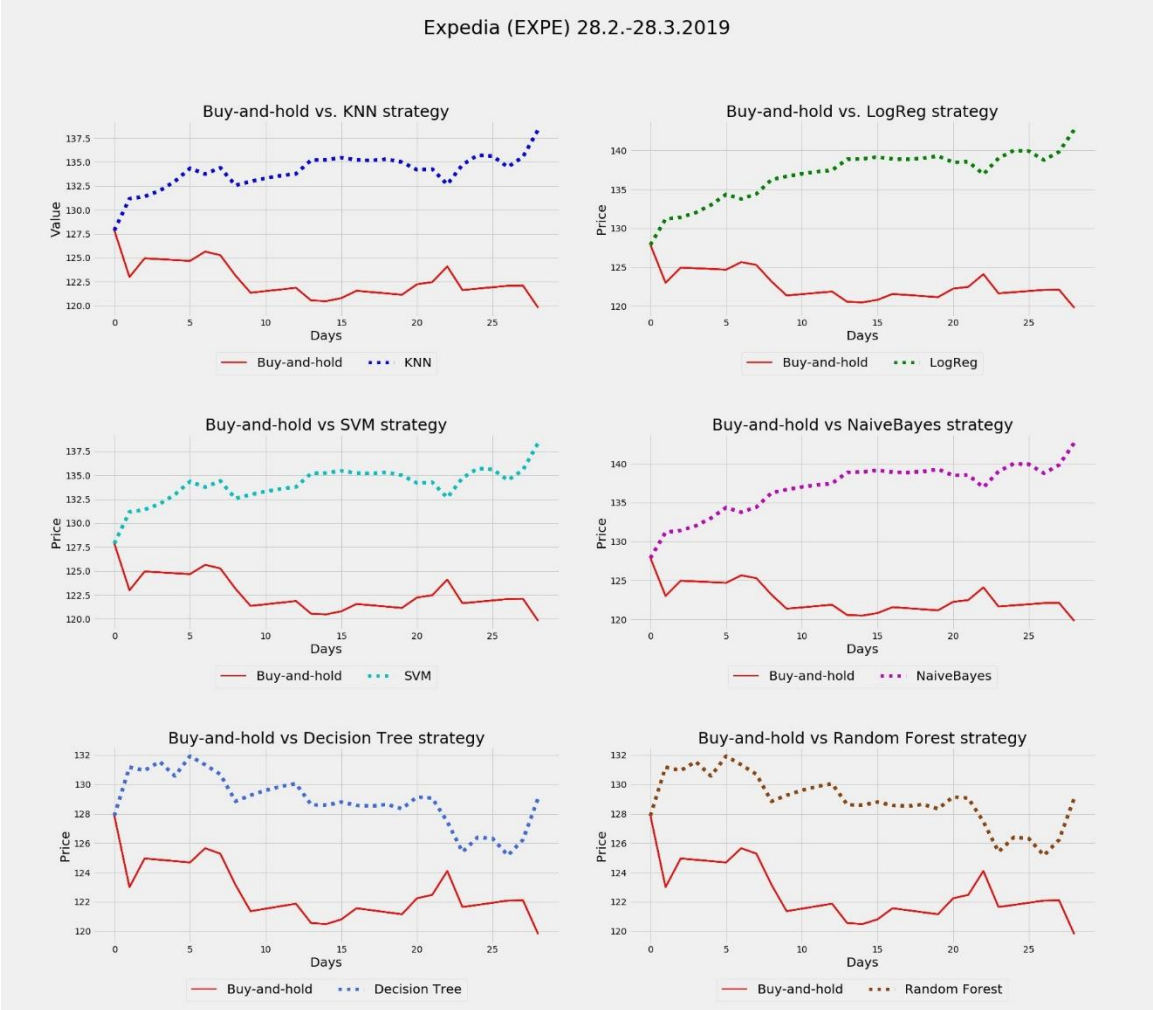
每个模型都使用 2016 年的原始推文进行训练。然后给出了每日建议：明天开盘时买入或卖出，收盘时卖出或买入。

看下图的策略结果：

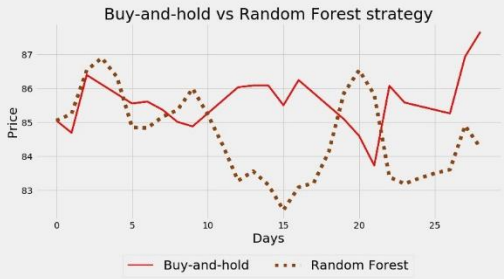

American Airlines (AAL) 28.2.-28.3.2019
Buy-and-hold vs. KNN strategy
Value
Days
Buy-and-hold
KNN
Buy-and-hold vs. LogReg strategy
Price
Days
Buy-and-hold
LogReg
Buy-and-hold vs SVM strategy
Price
Days
Buy-and-hold
SVM
Buy-and-hold vs NaiveBayes strategy
Price
Days
Buy-and-hold
NaiveBayes
Buy-and-hold vs Decision Tree strategy
Price
Days
Buy-and-hold
Decision Tree
Buy-and-hold vs Random Forest strategy
Price
Days
Buy-and-hold
Random Forest


Automatic Data Processing Inc (ADP) 28.2.-28.3.2019
Buy-and-hold vs. KNN strategy
Value
Days
Buy-and-hold
KNN
Buy-and-hold vs. LogReg strategy
Price
Days
Buy-and-hold
LogReg
Buy-and-hold vs SVM strategy
Price
Days
Buy-and-hold
SVM
Buy-and-hold vs NaiveBayes strategy
Price
Days
Buy-and-hold
NaiveBayes
Buy-and-hold vs Decision Tree strategy
Price
Days
Buy-and-hold
Decision Tree
Buy-and-hold vs Random Forest strategy
Price
Days
Buy-and-hold
Random Forest

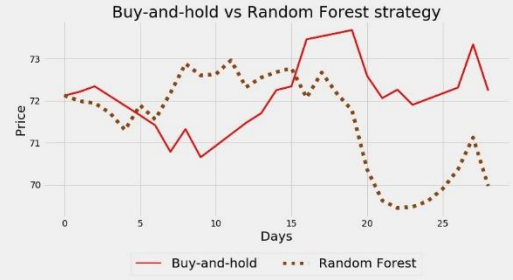

Cerner Corporation (CERN) 28.2.-28.3.2019
Buy-and-hold vs. KNN strategy
Value
Days
Buy-and-hold
KNN
Buy-and-hold vs. LogReg strategy
Price
Days
Buy-and-hold
LogReg
Buy-and-hold vs SVM strategy
Price
Days
Buy-and-hold
SVM
Buy-and-hold vs NaiveBayes strategy
Price
Days
Buy-and-hold
NaiveBayes
Buy-and-hold vs Decision Tree strategy
Price
Days
Buy-and-hold
Decision Tree
Buy-and-hold vs Random Forest strategy
Price
Days
Buy-and-hold
Random Forest

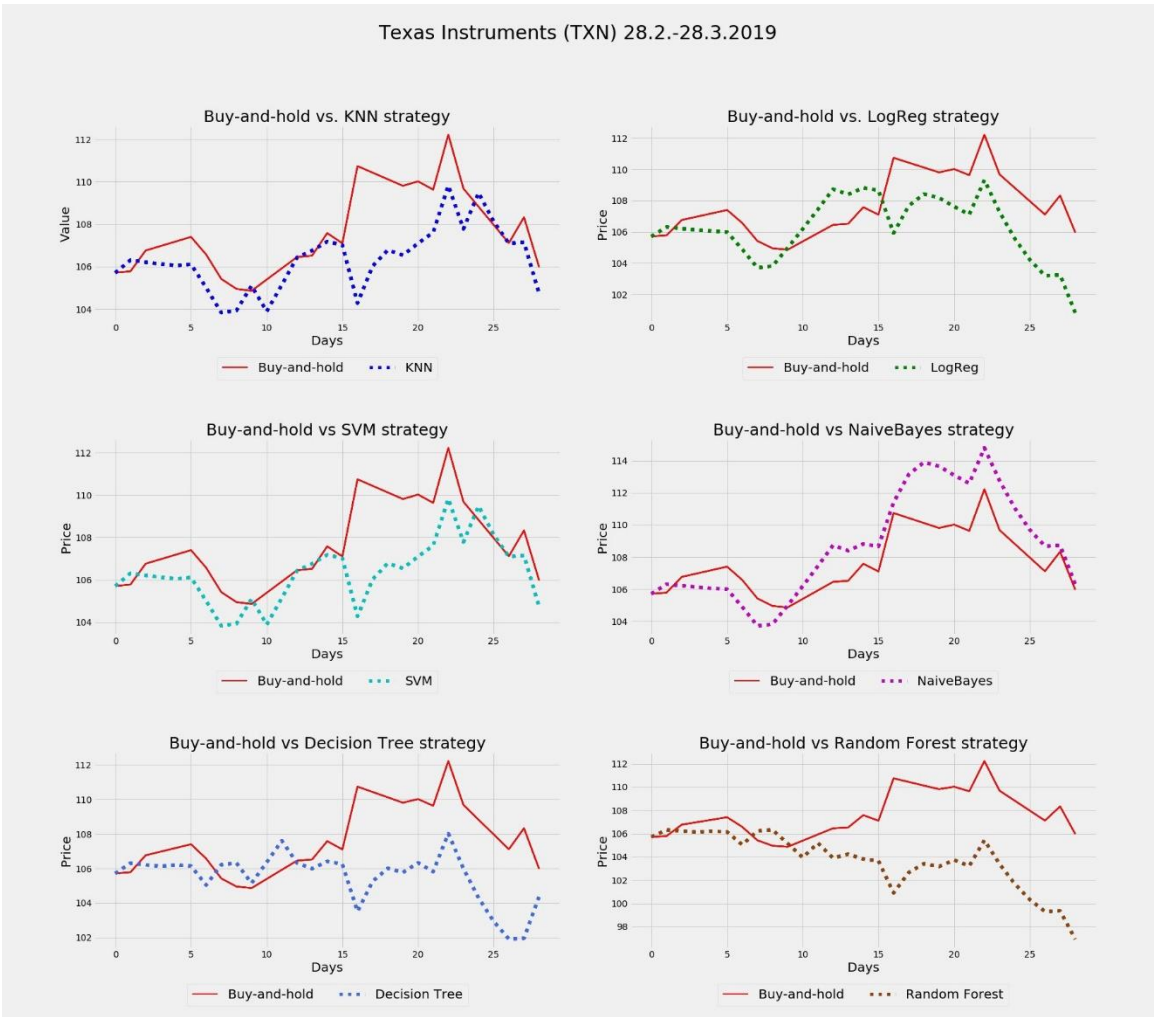

Expedia (EXPE) 28.2.-28.3.2019
Buy-and-hold vs. KNN strategy
Value
Days
Buy-and-hold
KNN
Buy-and-hold vs. LogReg strategy
Price
Days
Buy-and-hold
LogReg
Buy-and-hold vs SVM strategy
Price
Days
Buy-and-hold
SVM
Buy-and-hold vs NaiveBayes strategy
Price
Days
Buy-and-hold
NaiveBayes
Buy-and-hold vs Decision Tree strategy
Price
Days
Buy-and-hold
Decision Tree
Buy-and-hold vs Random Forest strategy
Price
Days
Buy-and-hold
Random Forest

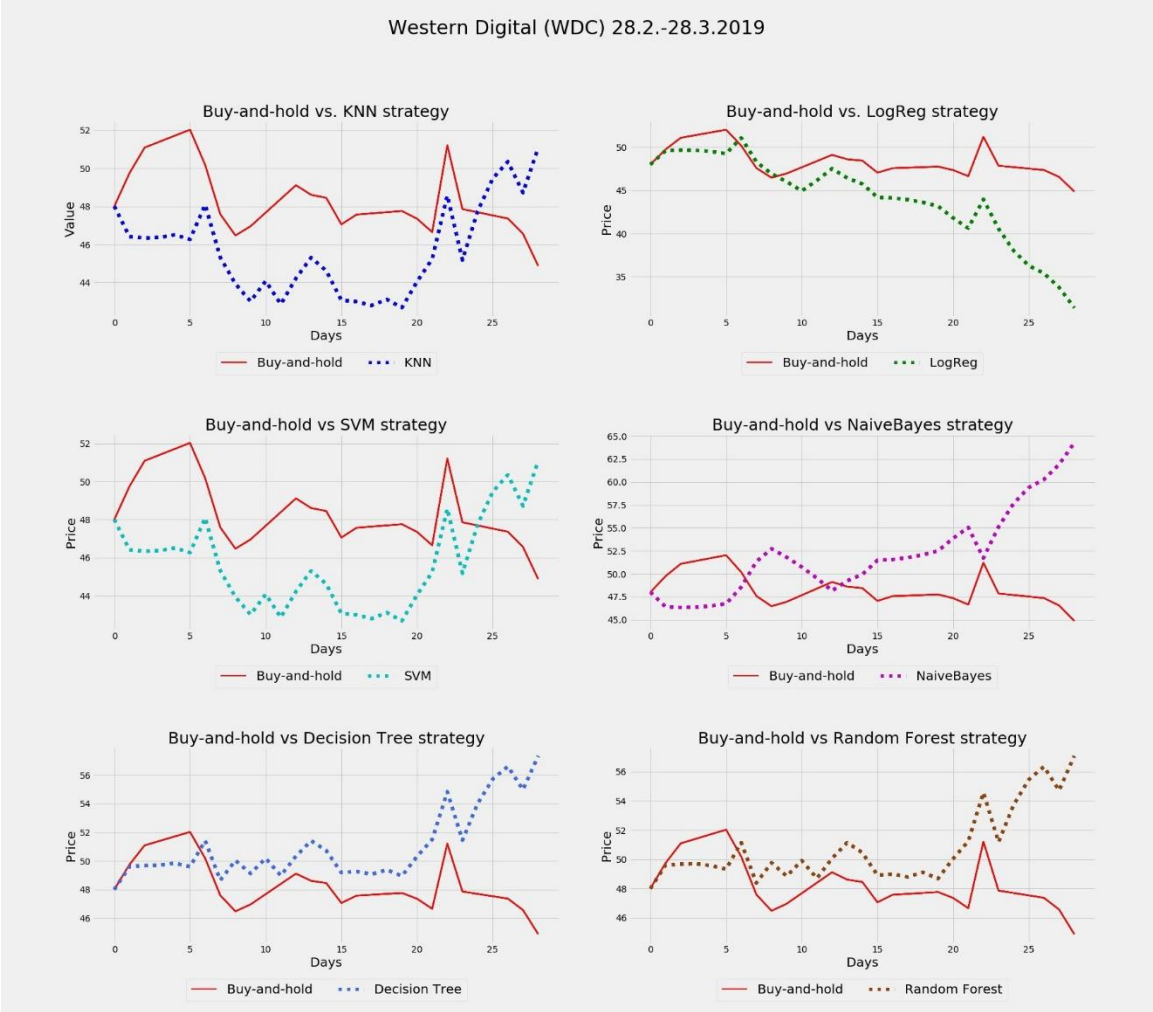
Fiserv (FISV) 28.2.-28.3.2019



T-Mobile US (TMUS) 28.2.-28.3.2019




Texas Instruments (TXN) 28.2.-28.3.2019
Buy-and-hold vs. KNN strategy
Value
Days
Buy-and-hold
KNN
Buy-and-hold vs. LogReg strategy
Price
Days
Buy-and-hold
LogReg
Buy-and-hold vs SVM strategy
Price
Days
Buy-and-hold
SVM
Buy-and-hold vs NaiveBayes strategy
Price
Days
Buy-and-hold
NaiveBayes
Buy-and-hold vs Decision Tree strategy
Price
Days
Buy-and-hold
Decision Tree
Buy-and-hold vs Random Forest strategy
Price
Days
Buy-and-hold
Random Forest



总结

总体而言，**基于情感的 twitter 策略在近 60%的模拟案例中击败了买入并持有策略。**

	Profit or loss from the 6 sentiment based models compared to buy-and-hold					
	LogReg	KNN	SVM linear	Naive Bayes	Decision Tree	Random Forest
AAL	profit	profit	profit	profit	profit	profit
ADP	loss	loss	loss	loss	loss	loss
CERN	profit	profit	profit	profit	loss	loss
EXPE	profit	profit	profit	profit	profit	profit
FISV	profit	loss	loss	loss	loss	loss
TMUS	profit	profit	profit	profit	loss	loss
TXN	loss	loss	loss	loss	loss	loss
WDC	profit	loss	profit	profit	profit	profit
Profit	6 out of 8	5 out of 8			3 out of 8	

除去两个表现最差的模型，决策树和随机森林，结果得到了进一步的改进。在三分之二的案例中，“买入并持有”不理想。

	Profit or loss from the 4 best performing models			
	LogReg	KNN	SVM linear	Naive Bayes
AAL	profit	profit	profit	profit
ADP	loss	loss	loss	loss
CERN	profit	profit	profit	profit
EXPE	profit	profit	profit	profit
FISV	profit	loss	loss	loss
TMUS	profit	profit	profit	profit
TXN	loss	loss	loss	loss
WDC	profit	loss	profit	profit
Profit:	6 out of 8	5 out of 8		

如果只遵循表现最好的模型 Logistic 回归，那么在 4 只股票中每 3 只股票就会盈利！

进一步完善模型思路

1、模型只有 75 天的数据用于训练和测试。如果情绪真的具有预测能力，那么从**更长的、甚至更近的时期添加更多数据，可能会显著改善结果。**

2、**为了使每周仅 5 天的股票数据与每周 7 天的 twitter 数据相吻合，需要对周末调整后的收盘价进行插值。**虽然考虑了特征工程，但周末创建的股票价格是人为的，可能会扭曲结果。**考虑到推文对周一股市走势的影响，或许周五到周日的推文应该以某种方式组合在一起。**

3、可以考虑将推特情绪的结果与其他技术结合使用，**比如 LSTM 神经网络进行时间序列分析，总是提前一天做出预测。**

4、尝试使用其他一些现成的模型，比如 **TextBlob**，而不是 VADER 来提取 tweet 情绪。或者更好的方法是，通过建立一个神经网络来训练你的情绪分类器，然后用你自己的数据来训练它，比如这里的数据；1.6mio 将每一行标记为 0=负，2=中性，4=正。

5、**时间对最终结果的影响有多大？**在模拟中，最终的 P/L 取决于周期的长度。在某些情况下，交易期越长，利润就会变成亏损，反之亦然。

6、**模拟中没有考虑交易成本。**至少在最终利润相当微薄的情况下，交易成本可以将利润变成亏损。

7、**能否在特定业务领域的特定股票中发现模式？**在这项分析中，美国航空和 Expedia 这两家旅游公司的股票收益最高。这仅仅是个巧合，还是某些企业的股票走势更容易引发推特情绪？