基于 twitter 情绪预测股市*

危嘉祺[1],

[1](北京大学 计算机科学技术研究所, 北京)
学号：1801213758
E-mail: jiaqi97@pku.edu.cn

摘　要: 行为金融学告诉我们，情绪可以深刻地影响个人行为和决策。因此，可以合理地假设公众情绪可以像新闻一样推动股市价格变化。本文对 Twitter 情绪与股价变化之间的关系进行了研究和分析。首先，我们对公开的 Twitter 数据集训练了一个基于双向 LSTM 的模型进行情感分析。然后，我们使用前几天 Twitter 情绪和前几天的股票涨跌数据来预测未来的股票走势。我们尝试了多个机器学习常用的分类模型来预测股价，包括：支持向量机（SVM）、朴素贝叶斯（Naive Bayes）、决策树（Decision Tree）、K 近邻（KNN）、随机森林（Random Forest）、逻辑回归（Logistic_Regression）。在这几种模型中都取得较好的效果，实现了超过 50%的准确度预测。

关键词: 股价预测;情感分析;Twitter

1 引言

股票市场预测引起了学术界和企业界的广泛关注。但真的可以预测股市吗？早期对股票市场预测的研究大多是基于随机游走理论和有效市场假说 EMH（Efficient Market Hypothesis）[1]。根据 EMH 股票市场价格主要是由新信息（即新闻）推动，而不是现在和过去的价格。由于新闻不可预测，股票市场价格将遵循随机游走模式，无法以超过 50%的准确度预测。

EMH 存在两个问题。首先，大量研究表明股票市场价格不遵循随机游走，确实可以在某种程度上预测，与 EMH 的基础假设不一致。其次，最近的研究表明，新闻可能无法预测，但可以从在线社交媒体（博客，Twitter 等）中提取非常早期的信息，以预测各种经济和商业指标的变化。例如，[2]展示了在线聊天活动如何预测图书销售。 [3]使用博客情绪评估来预测电影销售。 [4]使用概率潜在语义分析（PLSA）模型抽取博客的情感指数来预测未来产品销售。此外，谷歌搜索查询已被证明可以提供疾病感染率和消费者支出的早期指标[5]。 [6]研究了重量级财经新闻和股票价格变化之间的关系。 [7]展示了在 Twitter 上表达的与电影相关的公众情绪如何实际预测票房收入。

虽然新闻必然会影响股市价格，但公众情绪可能起着同样重要的作用。 我们从心理学研究中了解到，除了信息之外，情绪在人类决策中起着重要作用。 行为金融学进一步证明了财务决策受到情绪和情绪的显着驱动。 因此，可以合理地假设公众情绪可以像新闻一样推动股市价格变化。

本文对 Twitter 情绪与股价变化之间的关系进行了研究和分析，我们对公开的 Twitter 数据集使用一个基于双向 LSTM 的模型进行情感分析，分为三类：积极（positive）、中立（neutral）、消极（negative）。我们使用前几天 Twitter 情绪和前几天的股票涨跌数据来预测未来的股票走势。我们尝试了多个机器学习常用的分类模型来进行股价预测，包括：支持向量机（SVM）、朴素贝叶斯（Naive Bayes）、决策树（Decision Tree）、K 近邻（KNN）、

随机森林（Random Forest）、逻辑回归（Logistic_Regression）。在这几种模型中都取得较好的效果，实现了超过 50%的准确度预测。

2 研究框架

图 1 是我们基于 twitter 情绪预测股价的算法研究框架。首先我们需要获取两类数据：Twitter 数据以及股价历史数据，数据获取与数据预处理的过程我们将在第 3 章展开论述。然后对 Twitter 数据进行情感分析，情感分析其实是一个三分类任务，我们使用了一个基于双向 LSTM 的模型，具体细节我们将在第 4 章详细讨论。我们将前两天的公众情感值和股票涨跌数据作为特征，今日的涨跌作为预测标签，将所有数据划分为训练集和测试集。将训练集投入机器学习模型模型中进行训练，再在测试集进行回测验证。本次研究我们尝试了六种常见的机器学习分类模型，具体细节将在第 5 章给出。并且我们将在第 6 章给出我们的实验结果。

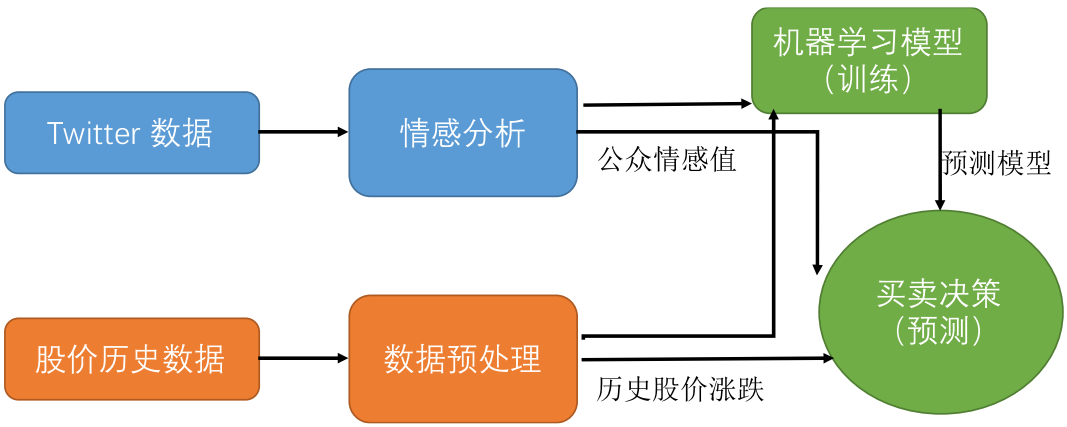


图 1 基于 twitter 情绪预测股价的研究框架

3 数据获取与预处理

3.1 Twitter数据

Twitter（推特）是国外的一个社交网络及微博客服务的网站，用户可以在 Twitter 上发表言论，该言论就是 Tweet。Twitter 的一项功能允许用户点击股票代码 Cashtags，比如$GOOG、$AAPL、
或$FB，看看用户在对应公司股票话题下的讨论。

首先我们假设 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。

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。可以从下面链接获取该数据：https://www.followthehashtag.com/datasets/nasdaq-100-companies-free-twitter-dataset/

表 1 为几个带有 cashtags 的 tweet 数据，分别取自$AAPL、$CSCO、$MSFT 三个 cashtags 的数据。每一行分别为该条 Tweet 的发表日期、内容和该条推文发布者的关注者数量。下面的数据是经过处理的，原数据包含的内容更多，下面三项为我们所需要的数据。

表 1 Twitter 数据示例

Date	Tweet Content	Following
2016/6/15	#APPLE TRIES to Limit #Google Incursions Onto Its Devices $AAPL $GOOG https://t.co/G8aecscj9S	346
2016/6/14	$CSCO Hot Tech Stocks To Watch Right Now: Cisco Systems, Inc. (CSCO), Advanced Micro ...: Hot Tech Stocks To... https://t.co/AtH730NdtN	6
2016/5/31	#Microsoft's new investment group targets cloud startups. Read more: https://t.co/pGH7HdjsGP $MSFT	2

由于情感分析模型计算量较大，计算资源有限，本文仅对四支股票进行了训练和预测分析，分别是：$AAPL、$CSCO、$MSFT 以及$INCT。

3.2 历史股价数据

使用 Python 开源库 yahoofinancials 下载 3 月 25 号到 6 月 15 号的股票数据。如表 2 所示，可以获取到每个工作日的股市开盘价、收盘价、最高价和最低价。我们增加一列涨跌幅数据（pct_change），计算方法为：（当日收盘价-前日收盘价）/前日收盘价。在第 2 章我们提到，我们使用前两天的 twitter 情绪和涨跌幅数据作为输入特征，当天的涨跌幅数据作为输出 label（涨为+1，跌为-1）。由于股市仅在工作日进行交易，周末没有股价数据，因此我们在训练和预测过程中，对于周一的前两天的数据，我们分别取得是前一周的周四和周五的数据。

表 2 历史股价数据示例

Date	Open	Close	High	Low	Pct_change
2016/3/28	106	105.1900024	106.1900024	105.0599976	-0.004542403
2016/3/29	104.8899994	107.6800003	107.7900009	104.8799973	0.023671431
2016/3/30	108.6500015	109.5599976	110.4199982	108.5999985	0.017459113

4 情感分析

4.1 模型设计

情感分析的任务可以定义为：给出一个推文，判断其表达的正面、负面、中立的情绪。我们采用基于深度学习的方法，设计了图 2 所示模型。我们使用了两层 Bidirectional LSTM，在 biLSTM 前加入了 Batch Normalization，Gaussian Noise 以及 dropout 层。在两层 biLSTM 后加入全连接 层，使用 Softmax 作为激活函数，输出最终的 class probabilities。

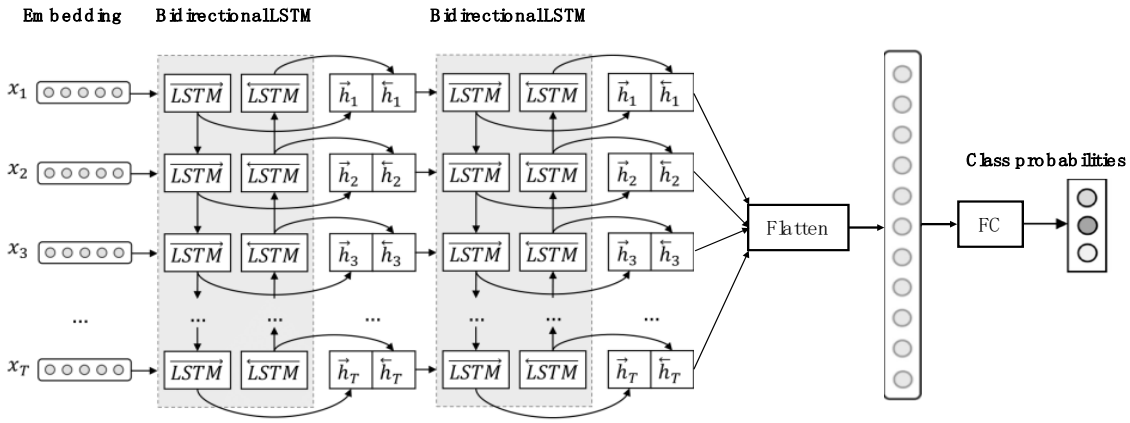


图 2 情感分类网络结构

程序语言：Python
深度学习框架：Keras
训练细节：将训练数据按 4:1 的比例切分为训练集和验证集，每条 twitter 的最大长度设置为 50，batch_size 设置为 128，learning rate 设为 0.001。

4.2 训练数据与数据预处理

a) 训练与测试数据集

数据集：SemEval-2017 Task4 Subtask
链接：http://alt.qcri.org/semeval2017/task4/index.php?id=results
训练集：twitter- 2016train-A.txt
测试集：twitter-2016test-A.txt

b) 数据预处理

使用开源库 ekphrasis（https://github.com/cbaziotis/ekphrasis.git）进行数据预处理，完成以下处理：

- 对文本中的 url，email，percent，money， phone number，time，date，number，user 等做归一化处理。
- 纠正拼写错误
- 词语切分
- 转换表情符号为对应的标签

c) 词向量

选择使用 GloVe 训练好的 word vectors
链接： http://nlp.stanford.edu/data/wordvecs/glove.twitter.27B.zip
向量维度：200d

d) 情感分类评测结果

评测指标：
Macro-Precision: 三类（正面、负面、中立）对应的 Precision 的平均值
Macro-Recall: 三类（正面、负面、中立）对应的 Recall 的平均值
Macro-F1: 三类（正面、负面、中立）对应的 F1 的平均值。其中 F1 定义如下：

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precison + recall}$$

表 3 情感分类评测结果

Macro-Precision	0.6482
Macro-Recall	0.6149
Macro-F1	0.6274

5 股价预测模型

这是一个二元分类任务，即结果要么是“买入”，要么是“卖出”。我们尝试了多个机器学习常用的分类模型，包括：支持向量机（SVM）、朴素贝叶斯（Naive Bayes）、决策树（Decision Tree）、K 近邻（KNN）、随机森林（Random Forest）、逻辑回归（Logistic_Regression）。

5.1 分类模型

a) 支持向量机

支持向量机（SVM，Support Vector Machine）是 Vapnik 根据统计学习理论提出的一种新的学习方法，它的最大特点是根据结构风险最小化准则，以最大化分类间隔构造最优分类超平面来提高学习机的泛化能力，较好地解决了了非线性，高维数，局部极小点等问题。对于分类问题，支持向量机算法根据区域中的样本计算该区域的决策曲面，由此确定 该区域中未知样本的类别.SVM 算法通俗的理解在二维上，就是找一分割线把两类分开。我们的目标是寻找一个超平面，使得离超平面比较近的点能有更大的间距。也就是我们不考虑所有的点都必须远离超平面，我们关心求得的超平面能够让所有点中离它最近的点具有最大间距。

实现方式：调用 sklearn 内置函数

```
from sklearn.multiclass import OneVsRestClassifier
from sklearn.svm import LinearSVC

clf_svm = OneVsRestClassifier(LinearSVC(random_state=0))
```

b) 朴素贝叶斯

贝叶斯（Bayes）分类算法是一类利用概率统计知识进行分类的算法，如朴素贝叶斯（Naive Bayes）算法。这些算法主要利用 Bayes 定理来预测一个未知类别的样本属于各个类别的可能性，选择其中可能性最大的一个类别作为该样本的最终类别。在 scikit-learn 中，一共有 3 个朴素贝叶斯的分类算法类。分别是 GaussianNB，MultinomialNB 和 BernoulliNB。其中 GaussianNB 就是先验为高斯分布的朴素贝叶斯，MultinomialNB 就是先验为多项式分布的朴素贝叶斯，而 BernoulliNB 就是先验为伯努利分布的朴素贝叶斯。这三个类适用的分类场景各不相同，一般来说，如果样本特征的分布大部分是连续值，使用 GaussianNB 会比较好。如果如果样本特征的分大部分是多元离散值，使用 MultinomialNB 比较合适。而如果样本特征是二元离散值或者很稀疏的多元离散值，应该使用 BernoulliNB。因此我们这里使用 BernoulliNB。

实现方法：

```
from sklearn.naive_bayes import BernoulliNB
clf = BernoulliNB(alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True)
```

c) 决策树

决策树又称为判定树，是运用于分类的一种树结构，其中的每个内部节点代表对某一属性的一次测试，每条边代表一个测试结果，叶节点代表某个类或类的分布。决策树的决策过程需要从决策树的根节点开始，待测数据与决策树中的特征节点进行比较，并按照比较结果选择选择下一比较分支，直到叶子节点作为最终的决策结果。

决策树学习过程：

- 特征选择：从训练数据的特征中选择一个特征作为当前节点的分裂标准（特征选择的标准不同产生了不同的特征决策树算法）。
- 决策树生成：根据所选特征评估标准，从上至下递归地生成子节点，直到数据集不可分则停止决策树停止声场。
- 剪枝：决策树容易过拟合，需要剪枝来缩小树的结构和规模（包括预剪枝和后剪枝）。

实现方法：

```
from sklearn import tree
clf_dt = tree.DecisionTreeClassifier()
```

d) K 近邻

k 近邻法（k-nearest neighbor, kNN）是一种基本分类与回归方法，其基本做法是：给定测试实例，基于某种距离度量找出训练集中与其最靠近的 k 个实例点，然后基于这 k 个最近邻的信息来进行预测。通常，在分类任务中可使用“投票法”，即选择这 k 个实例中出现最多的标记类别作为预测结果。

实现方法：

```
from sklearn.neighbors import KNeighborsClassifier
clf_knn = KNeighborsClassifier(n_neighbors=3)
```

e) 随机森林

随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习(Ensemble Learning)方法。首先使用 多棵决策树来单独预测，训练基分类器，然后由这些树的预测结果组合共同决定最后的结果。随机森林实际上是一种特殊的 bagging 方法，它将决策树用作 bagging 中的模型。首先，用 bootstrap 方法生成 m 个训练集，然后，对于每个训练集，构造一颗决策树，在节点找特征进行分裂的时候，并不是对所有特征找到能使得指标（如信息增益）最大的，而是在特征中随机抽取一部分特征，在抽到的特征中间找到最优解，应用于节点，进行分裂。随机森林的方法由于有了 bagging，也就是集成的思想在，实际上相当于对于样本和特征都进行了采样（如果把训练数据看成矩阵，就像实际中常见的那样，那么就是一个行和列都进行采样的过程），所以可以避免过拟合。

实现方法：

```
from sklearn.ensemble import RandomForestClassifier
clf_rf = RandomForestClassifier(max_depth=2, random_state=0)
```

f) 逻辑回归

逻辑回归虽然带有回归字样，但是逻辑回归属于分类算法。逻辑回归可以进行多分类操作，但由逻辑回归算法本身性质决定其更常用于二分类。逻辑回归是一种简单，常见的二分类模型，通过输入未知类别对象的属性特征序列得到对象所处的类别。由于 Y(x)是一个概率分布函数，因此对于二分类而言，离中心点的距离越远，其属于某一类的可能性就越大。对于常见二分类，逻辑回归通过一个区间分布进行划分，即如果 Y 值大于等于 0.5,则属于正样本，如果 Y 值小于 0.5，则属于负样本，这样就可以得到逻辑回归模型。

实现方法：

```
from sklearn.linear_model import LogisticRegression
clf_lr = LogisticRegression(C=1.0, penalty='l1', tol=0.01)
```

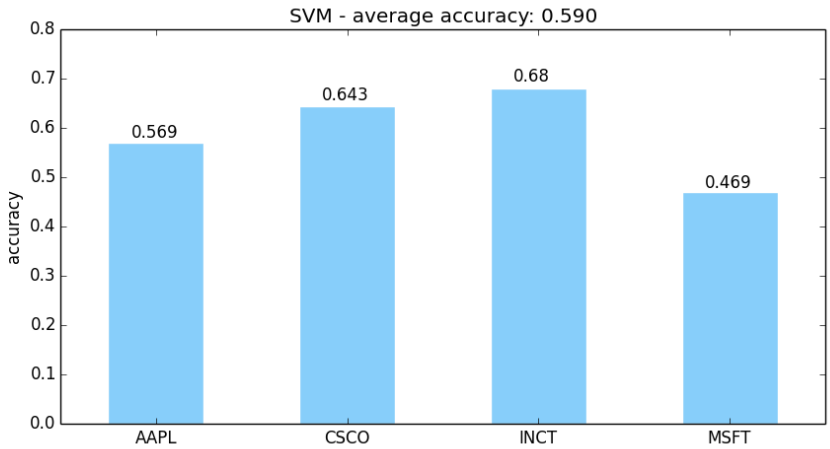
5.2 交叉验证

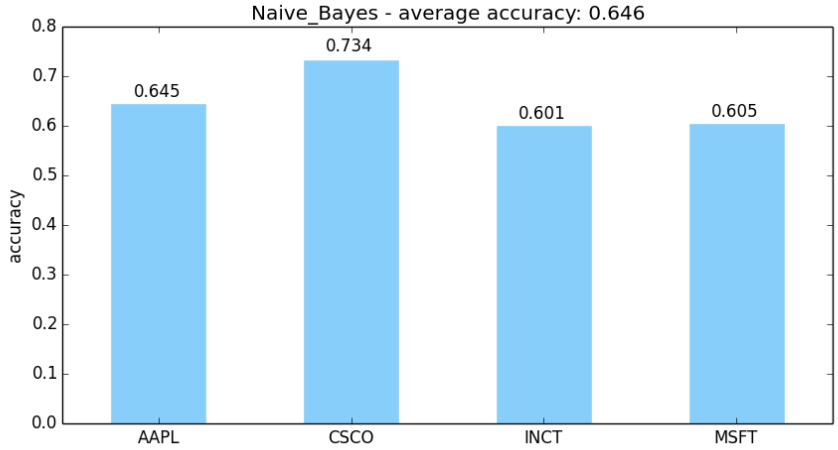
由于没有周末数据，我们的总数据量仅 53 天，数据量有限，仅使用 20%的数据（11 天）和 80%的训练数据（42 天）进行测试可能不够有代表性。为了避免训练/测试分割不完全随机的可能性，对数据进行交叉验证，这样得到每个算法精度更具代表性的结果。训练数据进一步分成 5 个子集，每个子集都与其他 4 个子集进行测试。

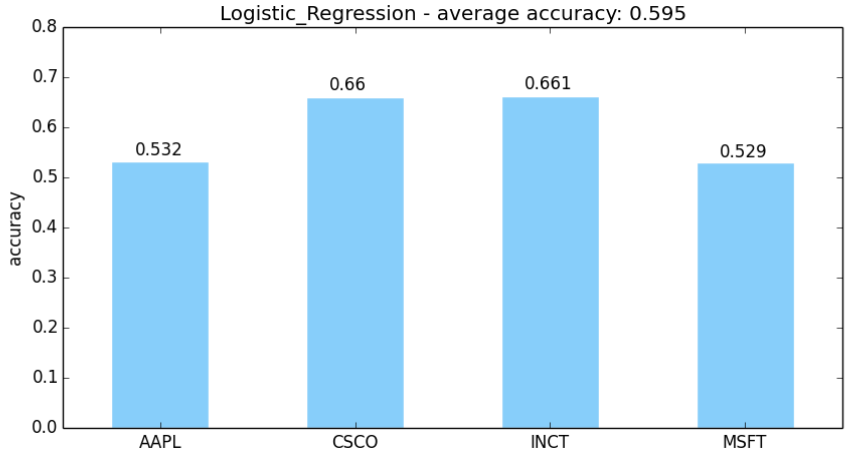
6 实验结果

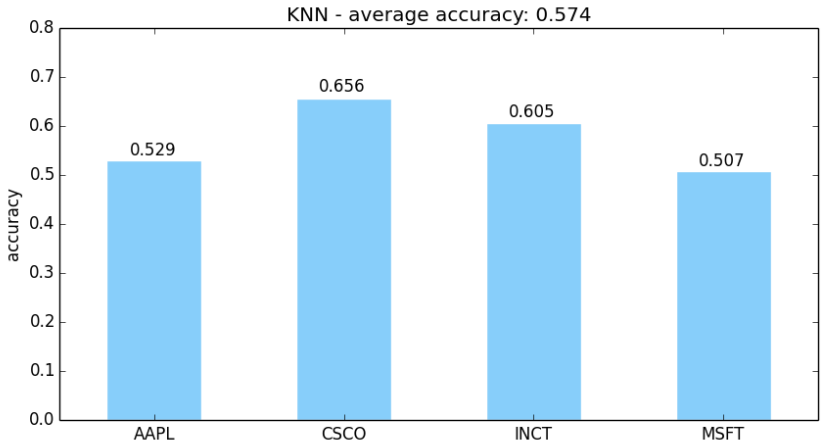
将 4 支股票分别通过 6 个二元分类器和 5 倍交叉验证后，结果挺可观。平均每个分类器的准确率都在 55%以上。这意味着，Twitter 上的情绪具有股价预测力，至少比随机决策强。准确率超过 50%在一定程度上证明了模型获得“非凡”收益的能力。更重要的是，对于许多股票，模型的准确性/预测能力在 60%以上！

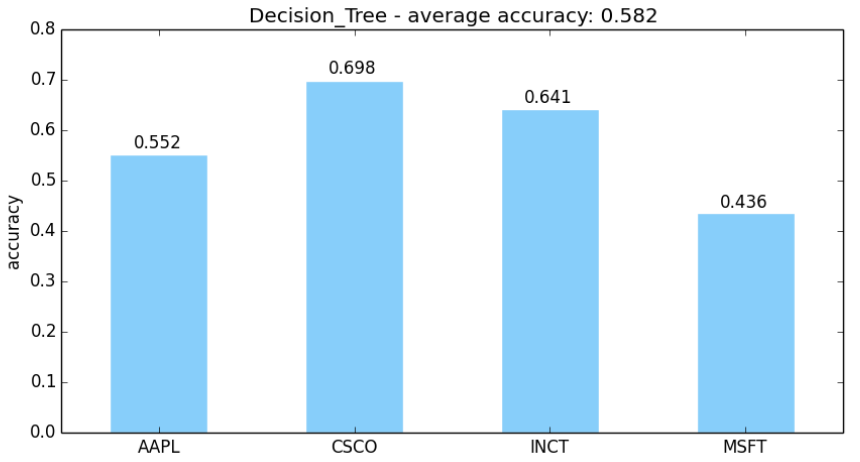
以下是所有分类器的准确率：

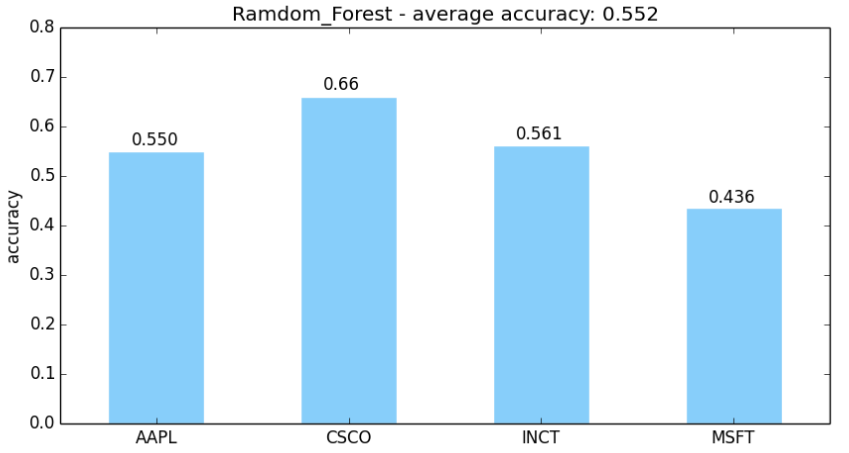



Naive_Bayes - average accuracy: 0.646
accuracy
0.645
0.734
0.601
0.605
AAPL
CSCO
INCT
MSFT


Logistic_Regression - average accuracy: 0.595
accuracy
0.532
0.66
0.661
0.529
AAPL
CSCO
INCT
MSFT


KNN - average accuracy: 0.574
accuracy
0.529
0.656
0.605
0.507
AAPL
CSCO
INCT
MSFT





7 总结与展望

综上，我们可以得出结论：基于 Twitter 情绪来预测股价变化是有一定的可行性的，能够提高收益率。预测性能最好的是朴素贝叶斯模型，其次是逻辑回归模型和 SVM 模型，然后是决策树和 KNN 模型，较差的是随机森林模型。六个模型都有一定的上升空间，如果进行进一步的调参会有更好的表现效果。 四支股票表现也是各异，其中表现最好的 CSCO 股票准确率能达到近 70%，而最差的 MSFT 略低于 50%，可能与获取的 Twitter 数据的分布不均匀与情感分类模型本身准确度并不是很高有关。总之，Twitter 情感分析对于股价预测是有一定增益的，未来可以有更多深入到的研究以提高预测到的准确率。

下面是我对于未来工作的展望：

- 考虑时序信息，由于现在的模型仅考虑了前两天的股价涨跌和 Twitter 情感，且不具备时序性。未来的研究工作可以尝试使用 RNN 进行时间序列分析。

- 提高情感分类模型性能。本次情感分类实验选择了基于双向 LSTM 的模型，模型较为简单， 由于 RNN 仅能捕捉时序信息，无法捕获到局部特征以及依赖关系等信息，因此后续工作可以结合 CNN 以及基于 attention 机制的 Transformer 设计新的模型。

References:

[1] Fama, E. F. (1965) *The Journal of Business* 38, 34–105.
[2] Gruhl, D, Guha, R, Kumar, R, Novak, J, & Tomkins, A. (2005) *The predictive power of online chatter*. (ACM, New York, NY, USA), pp. 78–87.
[3] Mishne, G & Glance, N. (2006) Predicting Movie Sales from Blogger Sentiment. AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs
[4] Liu, Y, Huang, X, An, A, & Yu, X. (2007) ARSA: a sentiment-aware model for predicting sales performance using blogs. (ACM, New York, NY, USA), pp. 607–614.
[5] Choi, H & Varian, H. (2009) Predicting the present with google trends., (Google), Technical report.
[6] Schumaker, R. P & Chen, H. (2009) *ACM Trans. Inf. Syst.* 27, 12:1– 12:19.
[7] S. Asur and B. A. Huberman 2010 Predicting the Future with Social *Media* arXiv:1003.5699v1