投资微博的情感分析

SENTIMENT ANALYSIS OF INVESTING WEIBO

目 录

CONTENTS

问题介绍

Introduction

数据预处理

Data Preprocessing

模型

Model

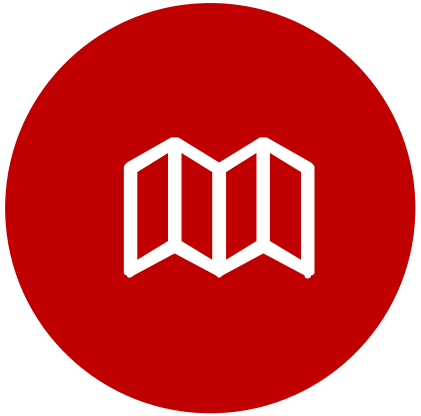
结论

Conclusion

Part 1　问题介绍

问题介绍

- 随着移动互联网时代的到来，人们更有意愿在网络上分享自己的生活、情感等，促使微博等社交网络中涌现出大量主观信息。
- 微博平台上就有许多投资者会将带有情感倾向（利好，利空， 中性）的投资言论以微博的形式发布于众。
- 从中可以挖掘用户的情感、观点，实现股票预测等任务为企业提供决策支持，因此基于投资微博进行情感分析具有重要的商业意义。



Part 2 数据预处理

问题介绍——初始数据

- 训练集为14844条、测试集为5000条
- 其中利空为4076条，利好为6125条，中性为4643条

Index	id	text	label
0	5405	【拉菲投资可以买入】近日拉菲价格大跌，上海2003年份拉菲从8660元跌至7740元，去年6月至今，高级葡萄…	利好
1	5409	想向大家推荐一篇分析性很强、概括面很广的文章．$中国银行{601988}$业市场化是大趋势，与非银行业的存…	中性
2	5468	【奇门测股】大盘接近2200点，今天六爻测盘代表官鬼的国家队进场护盘，但还不是买入时机，耐心等候7月7…	利好
3	5488	#开卷有益# 因此，多数散户都没有能力、时间和精神去选股，所以不要奢望跑赢大市，买指数基金即 可有利可…	中性
4	5489	#开卷有益# 葛拉汉《The Intelligent Investor》，曾被巴菲特誉为“最好的投资书”，他学…	中性
5	5493	昨日推荐的书籍《葛拉汉》（The intellegent investor）反应踊跃，@曹晓东_HUST、@理财新室 …	中性
6	5542	#中央大礼#虽自97后，香港已为中国一部分。但因制度上的不同，港人“北上”行商一直有所限制。直至2003年…	中性
7	5548	终于又一次降息了，市场流动资金是增加了，可是，这一举动，意味着什么？意味着6月的经济数据低于预期，意…	利空
8	5552	【丁未月可以入市】明天进入丁未月，之前也讲过这个月股市会上升，沪指能跌到2150点再入货是最理想的。大…	利好
9	237	没错，双赢。沪深两市市值高，但回报率不敢恭维，需要借鉴香港成熟的经验。	中性
10	260	$恒生銀行(000011)$ 香港股市主要指数周涨幅:恒生指数19800.64↗1.85%大市全周成交1849.2亿港元 国…	中性
11	262	$永亨銀行(000302)$ 目前，李嘉诚拥有多家蓝筹股公司，包括长江实业、和记黄埔、香港电灯及长江基建，…	中性
12	5580	今天大盘如意料之中暴跌，这说明中国经济的衰落，做空动能仍在释放中，不急于抄底，2012不远了。	中性
13	337	#信用评级的痛# 据数据显示，在2006年至2010年间，我国信用评级行业级别变动几乎均为级别上调，其中20…	中性

Index	id	text
0	173499	怪不得今日$方正证券{601901}$跌的如此惨啦，又被套了我的方正啊。
1	173501	$国农科技{000004}$ 在昨日T字涨停之后，今日继续涨停，集合竞价时间低开高走，之后股价一直拉平，只沿均线有小幅震荡…
2	173502	$泰亚股份{002517}$早盘走势比较平稳，午盘走势开始扭转，多方发动力量大幅拉涨，一路迅猛上升！
3	173503	$长城开发{000021}$今日股价回归20日均线上方，短线机会较大，但要看突破力度，谨慎操作吧
4	173504	$上海机场{600009}$慢慢修复了众均线及完全回补跳空缺口，近期走势以强劲为主，短线可紧密关注
5	173505	$亚太药业{002370}$ 今日大幅上涨，平开高走，早盘表现比较平稳，开盘一个小时之后开始大力拉涨，主力操作意愿强烈，…
6	173508	$中源协和{600645}$ 今日 生物医药股表现抢眼，呈现强势拉升的态势，午后涨幅再次扩大。$中源协和{600645}$的总体走…
7	173509	$独一味{002219}$ 上帝总算听到了散户们的祷告，今天总算大幅拉涨，前两日看涨的各位可是乐的笑呵呵了，一下子赚了差…
8	173510	$新北洋{002376}$新加入3D打印联盟。
9	173511	农业部、财政部等部门正在制定合作社专项扶持政策，资金大约为50亿元，相关个股：$辉隆股份{002556}$。
10	173514	$阳光城{000671}$在5日线上方小幅调整，量能呈萎缩状态，应该是正常的回调，无需担忧。//@机械师：$上海家化{600315}…
11	173515	$四环药业{000605}$ 今日大幅上涨，平开高走，呈上影阳线，上影线并不长，空方的力量还是比较弱的，不必太担心，成…
12	173516	$海峡股份{002320}$ 今日即被中散割掉，形成了主力、中散齐出局的局面，不适宜跟进了
13	173517	$航天机电{600151}$ 跌幅加快，先谨慎观望吧

数据预处理——去除标点和英文

Index	Type	Size	Value
0	unicode	1	【拉菲投资可以买入】近日拉菲价格大跌上海2003年份拉菲从8660元跌至7740元…
1	unicode	1	想向大家推荐一篇分析性很强概括面很广的文章中国银行{601988}业市场化是大趋势与非银行业的存贷活动更会有一场大搏奕http:tcnzWZ63R5
2	unicode	1	【奇门测股】大盘接近2200点今天六爻测盘代表官鬼的国家队进场护盘但还不是买入时机耐心等候7月7日前后吧
3	unicode	1	开卷有益因此多数散户都没有能力时间和精神去选股所以不要奢望跑赢大市买指数基金即可有利可获
4	unicode	1	开卷有益葛拉汉《TheIntelligentInvestor》曾被巴菲特誉为“最好的投资书”他学的是葛拉汉选股的思维但你不能学你可以从这本书取得的经是：要获得与 ...
5	unicode	1	昨日推荐的书籍《葛拉汉》Theintellegentinvestor反应踊跃曹晓东HUST理财新室等都指出受益于该书中的指数基金定投为让更多人读到好书受益以其中 ...
6	unicode	1	中央大礼虽自97后香港已为中国一部分但因制度上的不同港人“北上”行商一直有所…
7	unicode	1	终于又一次降息了市场流动资金是增加了可是这一举动意味着什么意味着6月的经…
8	unicode	1	【丁未月可以入市】明天进入丁未月之前也讲过这个月股市会上升沪指能跌到2150…
9	unicode	1	没错双赢沪深两市市值高但回报率不敢恭维需要借鉴香港成熟的经验
10	unicode	1	恆生銀行000011)香港股市主要指数周涨幅:恒生指数1980064↗185大市全周成交18492亿港元国企指数967962↗109红筹指数386293↗174 ...

去除前

Index	Type	Size	Value
0	unicode	1	拉菲投资可以买入近日拉菲价格大跌上海年份拉菲从元跌至元去年月至今高级葡萄…
1	unicode	1	想向大家推荐一篇分析性很强概括面很广的文章中国银行业市场化是大趋势与非银行业的存贷活动更会有一场大搏奕
2	unicode	1	奇门测股大盘接近点今天六爻测盘代表官鬼的国家队进场护盘但还不是买入时机耐心等候月日前后吧
3	unicode	1	开卷有益因此多数散户都没有能力时间和精神去选股所以不要奢望跑赢大市买指数基金即可有利可获
4	unicode	1	开卷有益葛拉汉曾被巴菲特誉为最好的投资书他学的是葛拉汉选股的思维但你不能…
5	unicode	1	昨日推荐的书籍葛拉汉反应踊跃曹晓东理财新室等都指出受益于该书中的指数基金…
6	unicode	1	中央大礼虽自后香港已为中国一部分但因制度上的不同港人北上行商一直有所限制…
7	unicode	1	终于又一次降息了市场流动资金是增加了可是这一举动意味着什么意味着月的经济…
8	unicode	1	丁未月可以入市明天进入丁未月之前也讲过这个月股市会上升沪指能跌到点再入货…
9	unicode	1	没错双赢沪深两市市值高但回报率不敢恭维需要借鉴香港成熟的经验
10	unicode	1	恆生銀行香港股市主要指数周涨幅恒生指数大市全周成交亿港元国企指数红筹指数标普香港创业板全周成交亿港元公用事业恒生工商分类恒生地产分类恒生金融分类

去除后

数据预处理——分词并去停用词

- Jieba分词
- 导入stopwords list

Index	Type	Size	Value
0	unicode	1	拉菲 投资 买入 近日 拉菲 价格 大跌 上海 年份 拉菲 从元 跌至 元 去年 高级 葡萄酒 指数 下跌 逾 投资回报 风险 极高 时期 高级 消费品 红酒 ...
1	unicode	1	想 推荐 一篇 分析 性 强 概括 面 广 文章 中国 银行业 市场化 趋势 与非 银行业 存贷 活动 更会 一场 大搏 奕
2	unicode	1	奇门 测股 大盘 接近 点 六爻 测盘 代表 官鬼 国家队 进场 护盘 买入 时机 耐心 等候 日前
3	unicode	1	开卷有益 散户 能力 时间 精神 选股 奢望 跑赢大市 买 指数 基金 即可 获
4	unicode	1	开卷有益 葛拉汉 巴菲特 誉为 投资 书 他学 葛拉汉 选股 思维 学 本书 经是 大市 同步 成绩 买入 指数 基金 跑赢大市 付出 努力 未必 成功 跑赢大 ...
5	unicode	1	推荐 书籍 葛拉汉 踊跃 曹晓东 理财 新室 指出 受益 该书 中 指数 基金定投 人读 好书 受益 投资 知识 请 关注 本微博 张宗永 转发 此条 微博 抽 ...
6	unicode	1	中央 大礼 虽自后 香港 中国 一部分 制度 港人 北上 行商 直至 首次 推出 消除 差别待遇 沪 深港 交易所 设立 合资公司 推出 买卖 基金 视为 中央 ...
7	unicode	1	终于 降息 市场 流动资金 增加 这一 举动 意味着 意味着 经济 数据 低于预期 意味着 经济的 不景气 意料之中 股市 反弹 力度 估计 弱 大盘 处于 探 ...
8	unicode	1	丁未 入市 明天 丁未 讲 股市 沪 指能 跌 点 入货 理想 股市 挑 家电 电子信息 燃气 供水 水力发电 地产 黄金 矿产 做 短线 丁未 戊申 一升 跌 ...
9	unicode	1	没错 双赢 沪深两市 市值 高 回报率 不敢恭维 借鉴 香港 成熟 经验
10	unicode	1	恆生 銀行 香港股市 指数 周 涨幅 恒生指数 市 全周 成交 亿港元 国企 指数 红筹 指数 标普 香港 创业板 全周 成交 亿港元 公用事业 恒生 工商 分 ...

数据预处理——词云


大盘
市场
行情
股价
公司
指数
反弹
上涨
板块
表现
关注
涨幅
股份
个股
投资者
持续
早盘
涨停
短期
后市
中国
股指
下跌
趋势
该股
行业
整体
走势


行情
公司
市场
大盘
股价
指数
反弹
板块
上涨
早盘
涨停
涨幅
中国
股份
个股
关注
表现
行业
投资者
股指
股票
该股
整体
走势
均线
持续
下跌
后市
震荡


行情
市场
大盘
股价
公司
反弹
指数
板块
股份
上涨
表现
中国
个股
涨停
早盘
投资者
股票
影响
关注
整体
走势
预期
下跌
趋势
回落
成交量
行业
震荡
持续
后市
均线
调整
股指
涨幅
业绩

数据预处理——词云

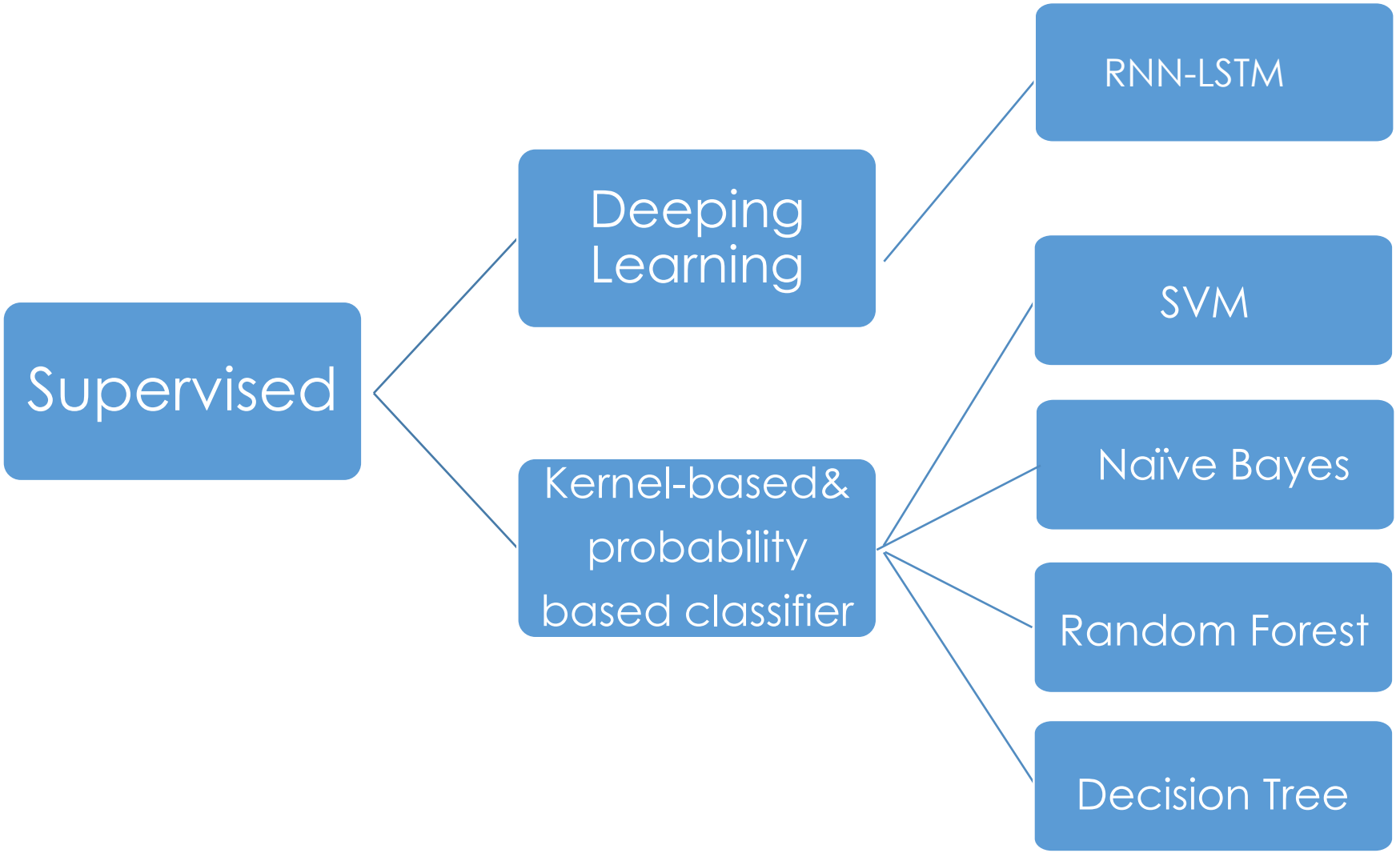
利空

利好

中性

Part 3 模型

分类算法——模型概览图


Supervised
Deeping Learning
Kernel-based& probability based classifier
RNN-LSTM
SVM
Naïve Bayes
Random Forest
Decision Tree

分类算法——机器学习(sklearn)

- 将带标签的数据按照7：3的比例分成训练集和验证集，并且保证训练集、验证集和原数据中的利空、利好、中性比例一致
- 分好词后的数据进行向量化，作为模型最终输入数据
- 调用sklearn中多种分类器训练模型

  （SVM、Naïve Bayes、Random Forest 、Decision Tree）
- 集成上述单模型，一种投票机制，也就是预测结果为综合概率最大的一类

分类算法——性能比较

- 调参之后，单模型和组合模型的准确率如下图所示
- 第一列是不去重的结果，第二列是去重的结果

  （重复数据有336条，可能原因是不去重保留重要变量并加大权重）

Classifier	Valid Acc	Acc(Duplicated)
Multinomial NB	0.6108431	0.5818015
SVC	0.6227889	0.6073070
Random Forest	0.6094647	0.5965074
Decision Tree	0.5074661	0.4956342
SVC + RF (soft)	0.6407723	0.6167279
SVC + NB (soft)	0.6253159	0.6036305

分类算法——深度学习(RNN尝试)

- RNN：循环神经网路，隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。
- LSTM：一种时间递归网络，解决之前普通RNN模型梯度爆炸或消失，无法对间隔时间很长的知识记忆的缺点。
- 我们参考了LSTM Networks for Sentiment Analysis中对于imdb（电影评论数据集）的情感分析代码框架。

*http://deeplearning.net/tutorial/lstm.html

分类算法——深度学习(RNN尝试)

- 构造文本数据的词库，将所有词转化为“索引”向量
- 在RNN中，每次训练只是一个句子，所以输入就是一个向量，但使用mini-batch之后会生成一个矩阵，多个句子并行训练
- 每个句子长度是不同的，为了便于并行矩阵计算，选定最长句子，最大化矩阵；句子词较少的，用0填充
- 输入数据，并利用AdaDelta算法自适应调整学习率
- 问题：运行较慢 内存不足

```
10156 训练集
4352 校验集
5000 测试集
开始训练过程
epoch 0 update 10 cost 1.10106462804
epoch 0 update 20 cost 1.13732147739
epoch 0 update 30 cost 1.03498713292
epoch 0 update 40 cost 1.07409218803
epoch 0 update 50 cost 1.06574599888
epoch 0 update 60 cost 1.08901374648
epoch 0 update 70 cost 1.11923250412
epoch 0 update 80 cost 1.07405449404
epoch 0 update 90 cost 1.06467937739
epoch 0 update 100 cost 1.06333878832
epoch 0 update 110 cost 1.08945700483
epoch 0 update 120 cost 1.06426594204
epoch 0 update 130 cost 1.06937011148
epoch 0 update 140 cost 1.07493689491
epoch 0 update 150 cost 1.08200038842
epoch 0 update 160 cost 1.12618507814
epoch 0 update 170 cost 1.0739761141
epoch 0 update 180 cost 1.06966452139
epoch 0 update 190 cost 1.06995272782
epoch 0 update 200 cost 1.10495598036
epoch 0 update 210 cost 1.08530512405
epoch 0 update 220 cost 1.11985697706
epoch 0 update 230 cost 1.0817387482
epoch 0 update 240 cost 1.07113229552
epoch 0 update 250 cost 1.0488352688
训练集错误率  0.590192989366 验证集错误率  0.578584558824
```

Part 4 结论

结果

- 根据分类器在验证集上的准确率表现选择最终模型
- 使用“soft”的投票方式集成SVC 和Random Forest两个模型
- 验证集上最终准确率为0.64，混淆矩阵如图所示（0：利空，1：利好，2：中性）

```
def voting_classify():
    clf1 = SVC(C=0.99, kernel = 'linear', probability=True)
    clf2 = MultinomialNB(alpha = 0.1)
    clf = VotingClassifier(estimators=[
        ('SVC',clf1),
        ('NB',clf2),
    ],
        voting='soft'
    )
    return clf
```

Confusion Matrix			
	预测值		
真实值	0	1	2
0	726	248	249
1	132	1453	253
2	255	463	675

不足与反思

- 停用词直接在网上找的文档，稍作修改，可能改成适合投资文本的效果会比较好
- 没有做特征选择，只是给出了每个词的权重，可能有很多噪声
- 模型的参数手工调试，函数运行太慢，应该还有提高空间
- 可以尝试用无监督算法跑出结果（时间来不及。。。）

Thank you !