

AAAI2020

在社交媒体平台上发布的投资信息对股票预测具有很高的价值。以往的研究大多将整体信息情绪视为预测指标，依赖浅层特征(词汇包、名词短语等)来确定投资意见信号。这些方法既没有捕捉到股票投资评论的时间敏感和目标感知特征，也没有考虑投资者可靠性的影响。在本研究中，我们深入分析了上市股票评论及其在股票走势预测中的应用。

Stock预测模型：

1. 时间敏感和目标感知的投资立场检测MFN，该模型通过集成多视角文本特征和金融领域扩展知识来学习每个评论的表示，从而提取看涨/看跌的投资意见。
2. 基于专家的动态立场聚合：识别每个评论的有效性，并通过结合基于专家的聚合意见信号增强股票走势预测。
3. 股票移动预测--门控循环单元（GRU）。

实验：xueqiu网站和guba网站，均是股票评论网站。

现有的大多数基于情绪分析的股票预测工作都将看涨情绪与整体消息情绪相一致，严重依赖单词级统计和传统分类方法。

缺点：

1. 它们在识别句子中多个意见与实体之间的匹配关系方面存在局限性。然而，作者很可能会赞成一个观点而反对另一个观点。以“Out SANY HEAVY at 8.95, bot back some CMBC this morning”这句话为例，预测指标为“Out”和“bot back”，代表对不同对象(两个大写单词)的相反观点(分别为看跌和看涨)。
2. 忽略了时间性。例如：I sold out that stock at 15.2 yesterday, but it may rise sharply over the next month。Sold out是yesterday（过去），rise是over the next month（未来），显然后者更有意义。有些评论只是过去的意见的总结与回顾，没有对未来的预测，哪怕对过去是积极的预测，也不应当认为是对未来看好的心态。
3. 以前的工作认为每个评论的价值是差不多的。尽管一些工作已经考虑了这一点，但也只是固定值或者不依赖时间的方法。用户和市场的信任值是动态变化的，而且每个投资者对每个领域的认知和价值也是不一样的。因此，建立一个动态更新的专家价值系统是很重要的。

基本结构：



以下对每个模块进行介绍，本文的细节描述的很详细。

1. 前提

时间参考模块-流程：

1. 日期-时间格式的文本首先通过与评论的发布时间进行比较来识别和分类。
2. 建立时间词典，识别与时间相关的单词和短语，例如last year等（现在都要建立字典了吗？）
3. 由于总结语义学通常指过去，而预测语义学指未来，因此建立了一个总结/预测词典来识别诸如“review”、“expect”等术语。

股票提取模块-流程：

股票名称抽取一个股票实体在回顾文本中可能有不同的提及。为了增强模型的目标感知能力，我们识别了股票提及对，即指代同一股票对象的文本中的实体提及对。我们的方法是基于词汇和语法匹配类型的规则。

规则条件和示例模式如下：对于基于词汇的模式，粗体子字符串被提取为对应股票的非规范提及。对于语法类型，“F”代表全称，有较长提及；“A”代表缩写，有较短提及。满足条件模式的单词和短语被提取为适当股票实体的不同提及。



观点三元组提取-过程：

结构化的表示方案TPO (T时间,P预测,O对象)

1. 首先，根据“金融词汇”和“时间参考和股票提取的过程”，识别出文本中所有的观点术语P、时间表达T和股票实体O。如果前面有负面项，则P将被转换为反向。
2. 对于每个P，我们找到离它最左边最近的一个作为这个P的引用时间。如果没有，时间部分将被替换为NOTIME。
3. 为了确定每个P所指向的对象，我们从P的左边和右边寻找最近的股票实体。如果一个对象O是多个意见术语P中最接近的实体，那么它将被分配给最接近的那个预测术语P。然后我们搜索第二最近的股票实体O作为其他TPO元组的对象候选。类似地，如果没有找到实体，对象部分将被标记为NOTAR。

举例：



1. 提取chase buy pick up等预测性词语
2. 提取下划线的时间性词语
3. 提取股票名称

结果：{(at present, do not chase, LeTV), (this is excellent time, buy, Gree), (will, pick up, SAIC Motor)}

观点三元组TPO的泛化-过程：

提取的TPO元组由于股票实体数量O庞大，以及意见P和时间T的同义词数量庞大，非常稀疏，这将影响我们模型的拟合能力。为了解决这个问题，我们赋予TPO中的每个术语其广义类。

例如，上述序列（CST/OST指需要关注的股票和其他股票）：

{(at present|T\_now, TRANSNEG|P\_bearish, LeTV|CST),

(this is excellent time|T\_now, buy|P\_bullish, Gree|OST),

(will|T\_future, pick up|P\_bullish, SAIC Motor|OST)}

1. Stance detection（立场检测）--MFN

目标：提取出正负向的评论。

形式化定义：

$$D=\left\{\left(r,s,u,t,y\right)\_{i}\right\}\_{i=1}^{N}$$

U是用户，t是发表时间（用于time-sensitive），r评论内容，s目标股票，y用户意见（积极、负面、中立）。目标是在D上建模g，对于任意的(r,s)，在给定的u和t下，预测积极性分数y。

* 1. Word-level表示

考虑原始输入$w\_{1:n}$：首先翻译成向量，然后通过对金融领域词汇的匹配，丰富词汇嵌入的相关知识属性，包括时间、倾向和否定词。换句话说，对于每个单词$w\_{k}$，最终输出的向量$e\_{k}$可以表示为：



其中，$e\_{f\_{i\_{k}}}$表示该词（第k个词）在第i个知识嵌入中获取到的向量。进入卷积网络。

* 1. TPO表示

以泛化后的TPO序列为输入，用于更好地建模评论中多个观点与股票实体之间的隐藏依赖和匹配关系。我们提出了一个词类对齐的多通道表示方法来联合学习每个特征空间的嵌入。给定一个广义的TPO三元组。步骤如下：

* + 1. 拼接词的嵌入和词对应的类的嵌入，将输出记为输出的注解：



* + 1. 现在，像2D图像那样，把TPO视为三通道的图像，如果一条评论有h个TPO，那么就可以组成以下的矩阵：



和2D图像略有不同的是，这里的e是一个$d\_{s}$维（即M\_TPO中的每个元素的维数），而图像里是一个数字，表示R/G/B的值。

也因此，卷积变成了多通道的卷积q，和word-level一样，进行卷积操作，并串接word-level的结果。

最后，我们将输出记为$s\_{o}$。

附：MFN方法图示：



立场预测：使用softmax和cross entropy即可。

1. 基于专家的立场聚合模块：

投资者对不同的股票有不同的专业知识，而且专业知识随着时间的推移而变化。为了区分高质量的投资意见，减少来自非专家的误导噪音，我们设计了一个动态的专家挖掘系统，并将其应用于每个股票。



算法过程：观察其预测和未来的预期是否一致，如果一致则增加C，否则增加I。最后，观察准确度A=C/(C+I)。取最高的50个专家。

细节：在这项工作中，评估间隔被设置为90天（这是通过实验得到的）。在此期间，投资者预测的可靠性相对稳定，发表的评论数量足以调查专业知识。对于股票预测，我们只根据作者的专家排名来估计每篇评论的质量。需要注意的是，在我们的框架中，专家排名系统会在每个交易日动态更新，以确保基于专家的聚合机制的有效性。

1. 独立的股票预测模块：

通过GRU来进行预测。输入是股票的历史数据，输出是情况。

GRU：





最终的预测：



Loss函数的选择：

1. 选择最后状态的预测值$h\_{t-1}$为损失
2. 将每个状态进行线性变换得到
3. 时间感知的注意力机制（非线性变换）



实验中三者都用到了，分别记为last、static、dynamic来进行比较。

实验一：立场检测模块

我们抓取了真实的股票评论数据集，其中包含了来自guba的6136128名用户撰写的165823986条评论(1/1/2010 ~ 6/10/2018)，以及来自xueqiu的218049名用户撰写的60522765条评论(9/11/2017 ~ 6/10/2018)，涉及3501只股票。然后，我们从大型语料库中随机选择20,000篇评论，并招募了三名标签人员将每个评论注释为三类(看涨/看跌/中性)。除了评论文本，标签上还提供了相关的信息，包括发表时间、作者和股票(我们根据评论发表在哪个首页来判断目标股票)。大多数标签都是最后决定的。Fleiss ' kappa (Fleiss 1971)的标签是0.74，表明一个相对较高的一致性（即这些标签人员标的是正确的）。



Baseline：TextCNN RCNN HAN Transformer

MFN消融实验（从第一行开始依次）：去掉TPO模块，去掉知识嵌入（采用纯词嵌入），去掉知识嵌入中的否定/倾向/时间嵌入

细节：

1. 我们为每个评论保留了最多120个单词，多出的截断，少的补0。
2. 验证集用于超参数的调优，并在测试集上进行最终比较。
3. 基线算法的参数按照论文初始化，然后进行调优，使性能达到最优。
4. 对于MFN，在两个编码器中，每个窗口（即卷积的尺寸）大小[2,3,4,5]的神经元数为60。
5. 词嵌入：采用word2vec训练词嵌入，维数为100。
6. 知识嵌入：每个知识嵌入都以50为维度进行随机初始化。
7. Adam优化器，mini-batch=32，lr=1e-3，早停法，dropout=0.5。

结果：



1. 没有word-level和TPO也比baseline好。
2. 以下是对编码器的理解：
	1. word-level编码器具有基于多视图局部交互的不同n-gram特征建模能力，
	2. TPO序列编码器更有助于增强全局交互和理解匹配关系。

这两个编码器是互补的，可以共同对整个模型的性能做出贡献。

1. 在扩展知识方面，这三个知识嵌入都是有用的，都可以超过TextCNN, TextCNN只是对单词嵌入序列进行卷积操作，说明了结合领域特定知识嵌入来增强股票评论文本表达的重要性。

实验二：股票预测模块

我们从公开可用的API tushare中检索价格数据，并将其与抓取的每日评论进行对齐。其中1/1/2010 ~ 31/12/2015的1456个交易日期用于训练，1/1/2016 ~ 30/7/2018的628个交易日期用于验证和测试，比例为1:1。我们回顾30天的历史信息来构建每个样本的输入向量。对中国股票市场12只知名个股(格力、建设银行、万科、上汽等)进行了实验，这些个股来自不同行业(金融、石油、汽车、电子信息技术等)。

细节：

1. GRU的隐藏状态维数为128
2. 虚拟上下文向量维数为50。
3. 采用Adam Optimizer对预测模型进行训练，
4. mini-batch=32，lr=1e-3，dropout=0.2

消融：



结果：



我们使用两个度量，精确度的标准度量(Acc)和Matthews相关系数(MCC)来评估单个股票预测，结果如图4所示。我们可以发现，结合专家投资意见可以在这两个指标上始终保持最佳表现。对于一些股票，准确率相当高，如苏宁67.78%，cypc 65.38%。

比较fs\_allr和fs\_pc，它证明了在大多数情况下，前者的性能更好。然而，有时使用所有的公共投资建议而不进行挑选会带来负面影响(例如左图的China Union)。我们认为，这是因为一些由非专家生成的低质量评论会在训练和测试阶段给预测模型引入大量的噪声和误导性的交易信号。此外，立场检测模块和基于专家的聚合模块对于最终的股票预测都是非常重要的。

实验三：loss的选择



结论：Dynamic效果最好。说明即使GRU有能力从输入序列中选择有用的信息，并通过其门机制将它们积累到最终状态，时间注意力机制仍然有助于提高该任务的性能。