基于微博用户评论与财经新闻分析的股市大盘预测（以上证指数为例）

选题意义

大盘是指沪市的“上证综合指数”和深市的“深证成分股指数”的股票。在本实验中，将以上证指数为例。大盘指数是运用统计学中的指数方法编制而成的，反映股市总体价格或某类股价变动和走势的指标。 上证综合指数是以上海证券交易所 挂牌上市的全部股票（包括A股和B股）为样本，以发行量 为权数（包括流通股本和非流通股本），以加权平均法计算，以1990年12月19日为基日，基日指数 定为100点的股价指数。

大部分时候，大盘好，个股也遍地开花，有些上涨趋势的个股一路向上；大盘不好，即便有个股暂时表现强势，后期补跌也是相当厉害（1月底至2月中的大幅下杀想必很多人都会记忆尤新），而对一些处于下降通道的个股更是雪上加霜。大盘和个股如水与舟，大盘强势，个股水涨船高，大盘走弱，个股逆水行舟，压力重重。成熟的投资者会把大盘的研判提升至首位，大盘的强弱与否直接关系到了自己的交易操作以及仓位控制，顺势而为才是盈利之道。因此，大盘指数不仅仅是宏观经济情况的晴雨表，更对股民有着重要意义。

值得注意的是，人类对于股市运行规律的认知，是一个极具挑战性的世界级难题。迄今为止，尚没有任何一种理论和方法能够令人信服并且经得起时间检验——2000年，美国著名经济学家罗伯特·席勒在《非理性繁荣》一书中指出：“我们应当牢记，股市定价并未形成一门完美的科学”；2013年，瑞典皇家科学院在授予罗伯特·席勒等人该年度诺贝尔经济学奖时指出：几乎没什么方法能准确预测未来几天或几周股市债市的走向，但也许可以通过研究对三年以上的价格进行预测。 但是，这并不妨碍我们做一些可能性的尝试。

我们知道，一些新闻事件往往会影响股市的涨跌，尤其是财经相关的新闻。因此，我们希望引入财经新闻的特征来对大盘的涨跌进行预测。此外，当一些热点的舆论事件发生的时候，用户或者说股民的态度可能对上证指数的涨跌有着影响。因此，我们尝试通过抽取新浪微博、财经新闻的数据，结合上证指数的涨跌信息，挖掘新浪微博和财经新闻与上证指数之间的相关性，并利用机器学习的模型，尝试建立预测模型，进而为股市投资者提供一定的参考信息。

本文主要的创新点在于：国内同时利用新浪微博和财经新闻来预测上证指数走势的研究尚未发现。一般的研究都基本基于财经新闻或者微博情绪，而我们将两者结合起来希望能提供更多有用的信息。

文献综述

预测上证指数，主要分为了两类的方法。第一种是利用各种经济学的指标来预测，包括上证指数本身、股票平均线、还有像长期利率等宏观经济指标，而第二类则是利用非经济学指标的其他信息，包括新闻、微博等。一直以来，更多的研究都是采用第一种方式，但是这些方法都忽视了除开经济学之外的更多影响因素，考虑到股市本质上是一个混沌系统，仅仅用经济学指标在信息论上存在着缺陷。

在使用经济学指标上，多年来，许多学者采用传统回归分析和时间序列方法对证券市场进行了预测和分析，通过证券价格的历史时间序列挖掘其变化趋势。然而这些传统预测方法是假设证券价格是呈线性趋势变化的，不能反映描述实际证券市场变化特点，预测结果可靠性不高。此外，还有一些学者尝试直接利用过去和现在和证券数据，通过支持向量机等方法来建立上证指数预测模型。【1】复旦大学的孙碧波提出移动证券有期不固定的移动平均线规则可以带来超额利润, 而持有期固定的移动平均线则是无用的 。这能在一定程度上预测上证指数，且时间越短，预测能力越强。【3】浙江大学的尚鹏岳等人认为宏观经济指标与上证只是之间有着协整关系。 其研究结果表明 1995 年 1 月到 2000 年 9 月这段时间内, 上证指数对长期利率、短期利率以及货币供应量的变化是敏感的, 但同国民生产总值、固定资产投资、全国物价指数的变化之间没有长期均衡的关系。【4】

在非经济学指标上，赖凯声等学者证明了微博情绪综合指数与同期的上证指数以及下一个交易日的上证指数之前明显存在的长期均衡关系。【2】香港科技大学的黄润鹏也认为，微博情绪信息反映的社会整体情绪倾向能够影响并预测股票市场整体价格走势的变化。【5】Bollen J等人也关注了Twitter 与股市的走势的关系，并证明了确实相关的。【7】这些研究为我们利用微博数据做上证指数预测模型提供了支持。

此外，有效市场假说认为，在强有效市场中，资产交易价格总是反映了所有的可得资讯 。股票市场将立即反应新的资讯，调整至新的价位。【6】而我们知道财经新闻可以是这种金融资讯的很好的体现，这也为我们使用新闻数据做上证指数预测提供了支持。

总体来说，利用非经济学因素对上证指数预测的研究，目前还比较少，目前找到的文献仅仅都关注了这些信息是否相关，但并没有构建一个实际的预测模型。因此，利用微博与新闻来预测上证指数的方法非常值得我们关注。

技术背景

在预测上证指数的时候，主要使用了word2vec embedding的方法来提取新闻和用户的微博内容。word2vec采用训练好的的中文word embedding vectors。这个word2vec模型是利用巨大的中文财经新闻来训练的。

然后我们使用了各种机器学习的分类模型对上证指数的涨跌做了预测，包括决策树、随机森林、KNN、逻辑回归、GradientBoosting、多层感知机等机器学习模型。这些机器学习模型都是很常见的分类算法，这里不再赘述他们的原理。

实验方法

- 数据：包括从2017-6-12到2018-5-11的财经新闻数据和对应日期的微博数据。我们利用2017-6-12到2018-4-17的数据作为训练集，利用2018-4-18到2018-5-11的数据作为测试集。这些数据均是通过爬虫得到的。
- 特征提取：在经过了大量的数据清洗之后，我们将当日的财经新闻数据利用word2vec的方法embedding到300维度的矩阵。同时对当日微博数据用word2vec的方法embedding到同样维度。然后对收集到的上证指数对应的第二日对应今日的涨跌作为目标项，涨高记作1，否则记作0。
- 训练数据：利用不同机器学习模型对提取了特征之后的微博或（和）新闻数据对涨跌进行训练。
- 预测涨跌：利用训练好的模型进行涨跌预测
- 评价模型：将预测和真实涨跌进行比较，评估模型。注意，如果是关注整个宏观经济状况，比如政府等模型使用者，应当关注accuracy，而如果是个人股民，则更希望关注他买入股票时机（预测涨的）是否是好时机（实际涨），此时应该关注precision这个指标，而不是accuracy。

实验结果

对于政府等关心宏观经济指标的，应该关注模型对所有涨跌情况的准确率，即accuracy

Accuracy	决策树	随机森林	**KNN**	多层感知机	**GBDT**
新闻	0.52	0.45	0.5	0.49	0.48
微博	0.46	0.47	0.5	0.51	0.41
新闻+微博	0.43	0.45	/	0.50	**0.54**

对于股民等关心买入时机的，应该只关注模型对涨的情况，即precision

Accuracy	决策树	随机森林	**KNN**	多层感知机	**GBDT**
新闻	0.53	0.39	0.5	0.19	0.47
微博	0.46	0.45	0.5	0.32	0.34
新闻+微博	0.42	0.40	/	0.26	**0.60**

上述/表示f1-score为0的情形，即模型在此时并不具有实际的作用。

结果分析

横向对比，我们发现，在accuracy任务和precision任务中，GBDT（*Gradient Boosting* ）都有着最好的效果。这可能是因为GBDT在训练中，通过改变训练样本的权重（增加分错样本的权重，减小分队样本的的权重，学习了多个分类器，并将这些分类器线性组合，提高了分类器性能。 boosting算法开始时，为每一个样本赋上一个权重值。在每一步训练中得到的模型，会使得数据点的估计有对有错，在每一步结束后，增加分错的点的权重，减少分对的点的权重，这样使得某些点如果老是被分错，那么就会被“严重关注”，也就被赋上一个很高的权重。然后等进行了N次迭代（由用户指定），将会得到N个简单的分类器（basic learner），然后我们将它们组合起来（比如说可以对它们进行加权、或者让它们进行投票等），得到一个最终的模型。 这可能是GBDT效果号的原因。

纵向对比，我们发现当新闻和微博的信息都利用的时候，我们的预测模型有着较好的效果，说明我们的创新工作-同时使用微博与新闻确实是有意义的。

accuracy和precision对比，我们发现precision在更多的时候表现要较好，这说明我们的预测模型更适合做个人股民的投资的辅助，而不适合政府对宏观经济数据的把控。

参考文献

【1】陈海英. 基于支持向量机的上证指数预测和分析. 计算机仿真. (2013)

【2】赖凯声，陈浩，钱卫宁，周傲英. 微博情绪与中国股市：基于协整分析. 系统科学和数学. (2014,5)

【3】孙碧波. 移动平均线有用吗？数量经济技术经济研究.（2005）

【4】尚鹏岳，李胜宏. 上证指数与宏观经济指标协整关系的实证分析. 预测 (2002, 4)

【5】黄润鹏，左文明，毕凌燕. 基于微博情绪信息的股票市场预测. 管理工程学报. (2015, 1)

【6】Fama EF． Efficient capital markets: a review of theory and empirical work［J］． The Journal of Finance，1970，25(2): 383

【7】 Bollen J，Mao H，Zeng X． Twitter mood predicts the stock market［J］． Journal of Computer Science，2010，2(1): 1 ～8.