提要

越来越多的推特用户使得研究正在发生的事情成为一种有价值的信息资源。用户经常使用推特来记录生活事件。在这里，我们只关心金融社区（financial community）的问题。这篇论文的重点是通过情感分析来检测在推特网络中事情的受欢迎程度，检测事件的流行度。这是一件非常重要的任务，因为推文通常很杂，这项工作的目的是要过滤掉所有嘈杂的的推文，只分析影响金融市场的推文。更具体地说，是研究构成道琼斯指数的30家公司。在本文提出了一种从推特金融社区过滤收集的推文，对推文进行情感分析最终检测到公司运营中的重要信息的方法。

第一节 介绍

由于变化的速度以及金融市场复杂性的增加，有必要创建技术工具，帮助投资者正确的运用他们的资产，以获得巨大的利润。由于股票市场是投资者的一种盈利形式，而这种类型的投资有很高的相关风险，使其无法保证一定的利润。对投资者来说，价格预测以及投资或退出股票市场的时机是一个很热门的话题。

为了增加利润和减少风险，投资者已经知道并应用了几种技术。其中最常用的技术分析是基于假设过往的模式与每种资产的价格行为有关，在未来可能会重复发生。另一方面，基本面分析也是投资者使用的另一种分析方法，它着眼于宏观指标或公司的资产负债表，它还通过基本分析的结果来确定资产是否被高估或低估。

某只股票的较大或较低的供给需求可能受到价格的历史行为、公司未来股票发行前景甚至计划的影响。通过在博客，社交网络上发表的新闻。某某某的研究表明：信息方面的文本的情感方面的新闻可能会影响金融市场业务量和股票价格的波动水平。对特定新闻的情感分析是通过一种叫做情绪分析的技术进行的，这对于了解新闻对于金融市场的影响是积极地还是消极的是至关重要的。

今天，互联网被用来寻找信息，也用来分享信息、分享知识，也可以作为一种渠道。在线社交网络每天都是全球数百万人生活中的一部分，多年来成为了一个重要的交流平台，汇集了各种信息，包括用户在分享新闻信息时表达的观点和感受。社交网络吸引了许多研究人员的注意力，他们的目标是将各种出版物的内容与现实生活中的事件联系起来。考虑到本文关注的焦点是股票市场我们只关注能改变特定股票演变的重大事件，我们只关心消息是正面的还是负面的。推特是一个非常嘈杂的新闻来源，这项工作为事件的流行提供了一个基础，它必须处理大量的数据和众多的社交网络的杂音。推特是一个在线社交网络，允许用户从其他的链接发送和接收个人更新。推文是一种针对事件流行的系统，我们主要关注构成道琼斯指数的公司的日常事件该方法由几个阶段组成：从推特金融社区的构建，通过推文的收集，对推文内容的分析，直到最后一步的事件检测和分析判断人们喜欢或不喜欢。

举几个例子：产品发布、公司日常的特殊事件、竞选活动等。本文提出了一种关于性质的原始研究。在这里，我们研究了构成道琼斯指数的30家公司直接相关的事件以及每个公司的金融用户的意见。首先，在推特上定义一个金融社区的算法，有四个主要贡献：对每个用户提取所有的推文、然后推特过滤算法使我们能够更恰当的将推文重新定位到金融市场。第三个贡献是创建了一个简单的情绪分析工具和与其他三个已知的情绪分析算法的结合。通过四种算法，我们分别以积极、消极和中立的分数对推文进行评价。对于每一个公司来说，算法创造的每日得分与当天的推文的得分之和有关。本文的最后一个贡献是手动检测具有较大正峰和负峰的事件，并直接与积极和消极影响金融社区的特殊事件相关。

本文通过社交网络推特提供了一种新的事件流行方法的详细讨论。本文的结构如下：第二节阐述了开发工作背后的理论，即一些与市场分析相关的概念，作为技术分析和基础分析。在这一节中，也有一些工具是情感分析的一部分，现在有很多研究方法可以提取出社交网络上发表的新闻和评论的情绪。最后，本节介绍了一些关于社交网络推特的相关文献，是关于它们对市场的影响以及在推特的社交网络事件监测领域的一些作品。第三节说明了每个模块的系统架构和各自的功能。第四节提出了评价发展战略的验证过程，提出了三个案例研究，以证明该方法的受欢迎程度。第五节概述了所提供的文件，提供了各自的结论和可能的未来工作的建议。

1. 2.1市场分析

法玛（1970）提出了有效市场的概念，随后进行了几项研究试图验证这一理论。这一理论是金融理论中最重要、最具争议性的问题之一。根据这个理论，当金融产品的价格迅速反映市场上可用信息的任何变化时，阻止了不寻常收益的实现时市场被认为是有效的。《随机漫步理论》（2006）认为，你不能从股票短期的变化和过去走势预测未来市场走势，价格走势不可预测，这一理论的创造者莫里斯.肯德尔也被定义为“随机漫步者”。根据豪根（2001）的研究，有效市场的价格行为是投资者理性行为的产物，而在低效率的市场中，价格行为是由利益相关者的情感和心理状态所导致的。

与此同时，一些投资者认为，你可以通过基本面分析或技术分析来击败市场。基础分析从其经济和财务信息评估业务。从它的数字知道一个公司的值，你可以将此值与市场分配给它的值进行比较。这些数字来自于整体经济，特定产业的部分或最典型的—来自公司本身。诸如通货膨胀、失业、股本回报率（ROE）、债务水平和个人价格比率（PE）等数据都可以在决定股票价格上起作用。

分析股票市场和股票价格未来演变的另一种方法是技术分析。90年来，这种市场分析方法是由著名的道琼斯工业平均指数（DJIA）创始人之一、《华尔街日报》创始人查尔斯.亨利.陶氏提出的。他的理论基于三个主要假设：

价格是所有市场力量的综合反映。在任何时候，所有的市场信息及其优势都反映在价格上；价格的走势可以被识别并转化为利润机会：价格变动在历史上是重复的。

如上所述，许多投资者一直在研究一种利用基本分析和技术分析的各种算法来预测未来价格的方法。这些 工具被专业人士或业余投机者用来分析一些金融资产的价格变动。影响股票价格的主要因素是需求和供给。但影响股价的还有其他方面。在信息时代，新闻可以传播到世界各地，有时比发生的速度更快。然而，新闻并不是传播影响金融市场的信息的唯一手段。每天侵入社会网络的的评论和出版物也允许我们提取出关于信息来源者的动机类型的非常具体的信息。这些动机可能有隐含的消极或积极的感觉。今天，可以根据互联网上更具体地在社交网络上找到的信息来创建新的指标。本文的目的是利用来自社会网络的信息，创建可用于检测股票市场企业的新指标。

2.2对文本分类的情感分析

互联网的广泛扩张产生了不同的信息，而这些信息问问隐含观点。某某某和某某某提到，这些观点非常重要，无论他们想在哪里做决定，人们都想听听别人的意见。这不仅适用于人，也适用于那些已经看到客户对他们的产品和服务有意见的组织，作为组织的附加价值。

对情绪或观点挖掘的分析是对文章中表达的观点、情绪和情感的计算研究。在社会媒体分析的节目中，由公司，营销人员、政治评论员，某某某等提出这一技术被广泛研究作为一种工具。而弗朗西斯科提出了另一种新的方法，利用无监督的依赖性分析来预测非正式文本中的情绪。作者在不经过训练的情况下，实现了一种基于情感传播的算法。结果证实了系统的竞争性和可靠性。

为了探索意见挖掘的领域，某某某和某某某提出了一个关于这个领域的信息融合的调查。作者对最流行的观点挖掘技术进行了调查，定义了信息融合领域。

信息融合是研究将不同来源的信息转换成单一的连贯表示的有效方法的领域 本文提出了一种指导融合过程的框架并回顾了一些在意见挖掘的背景下成功地实现了意见挖掘系统和信息融合技术的研究。

利用数据挖掘技术来预测金融市场已经在许多出版物中得到了广泛的的研究。某某某和某某某（2009）提出了一项研究，目的是通过使用SVR算法在金融新闻文章中应用文本挖掘技术来发现标准普尔500指数所列股票的实际价格。用户发布的在线文档和文本的激增导致了情绪分析领域最近的增长以及它与金融市场的关系。最近，某某某和某某某（2014）发表了一篇文章，目的是为了评估利用数据挖掘方法在日内交易中预测股票收益的有效性。另一篇论文证明了金融新闻文章对股票价格回报的影响，是由某某某等人（2014）出版的。本文分析了情绪维度对新闻的影响。在第一阶段，作者实现了一个通用股价预测框架。在第二阶段，他们使用两种不同字典构造情感维度。作者评估模型的预测准确性和实证比较他们的性能在不同的市场水平。结果表明情绪分析确实能帮助提高预测精度。但是，仅仅关注消极和积极地维度不能带来有用的的预测，最终作者得出结论，使用不同情感字典的模型之间存在细微的差别。

几项研究概述了情感分类中使用的一些技巧词汇资源的情感分析已经引起了计算语言学家的极大兴趣。布莱德利和朗（1999）在英语词汇的情感规范下发布了新的词汇。尼尔森（2011）探索了新推特的应用，利用AFINN词汇。Jain和Nemade（2010）在积极和消极的分类中标注了英语单词的列表，释放了“意见查找器”词典。词汇资源的开发利用了实力评估的方法。在2012年，又用到了情感力量。Esuli和Sebastian（2006）和后来的Baccianella将著名的Wordnet词汇数据库推广到一些概念集，创造了多愁善感的网络。

2.3 推特

因为这些年来发展壮大的社交网络，一些研究注意力集中于情感分析工具在出版物、用户发布的许可、评论和文章中的应用。社交网络在互联网的大宇宙中脱颖而出。因为它加速沟通而且允许任何一个人成为一个内容生产商。某等（2013）表明大多数金融市场专业人士和客户使用社交网络是因为一些专业原因。这是一个很有启发性的趋势——社交网络对市场有显著的影响力。今天，推特推特对此类型的工作是一种重要的信息来源。因此，推特已经赢得了越来越多的空间，让一个忠实的个人和他的关系群体的行为图景得以实现。某等（2011）研究了从大规模的推特衍生的集体情绪状态的测量是否与推特反馈与道琼斯工业平均指数（DJIA）的价值相关。他们通过两种情绪追踪工具来分析每日推特的文本内容，即“观点者”，衡量情绪状态的积极与消极情绪和谷歌的情绪状态（GPOMS），以6个维度来衡量情绪（平静、警觉、自信、重要、善良和快乐）

推特用户的意图是不同的。有些人仅仅用推特来谈论他们的日常活动，其他人则是出于职业原因使用它，而其他人则用它来传播恶意内容。为了了解用户在推特上的影响力，某等（2014）发表了一项研究，其中包括在推特上形成一个金融社区，在这个社区里，用户在金融市场上有共同的利益。结果表明，社区气候与股票价格变动之间的强烈相互依赖可以创造出金融共同体。本研究可以得出结论，金融共同体各节点所产生的情绪在一致的市场回报和市场波动中具有预测力。

推特发布推文有140个字符的限制，有时很多用户无法用最好的方式表达自己，所以他们使用表情符号。因此，使用较少的字母，用户可以表达诸如快乐、厌恶、愤怒、害羞等情绪。在接下来的文章中，作者使用特定的表情符号来形成情绪分类训练集。（某某2009）提出了一种自动分类推特信息的方法，这些消息在查询词中是正的或负的。训练数据包括带有表情符号的推特信息，这些表情符号被用作嘈杂的标签。该工作的方法是使用不同的机器学习分类器和资源提取器。

关于网络论坛、博客和推特的广泛讨论和研究是另一种形式的政治辩论。一些研究人员已经承认了政治类博客的质量，而另一些人则怀疑博客聚合和传递信息的能力。一些研究案例发现，在线信息已经相当成功地成为选举成功的指示器。由（某某某2012）出版的一篇论文描述了在2012年美国大选中对总统候选人的公众情绪的实时分析系统，他们在这一分析中表达了这一观点，他们试图探索推特是否能深入的了解工众舆论的变化和迹象。在他们的系统中使用的情绪模型的设计是基于这样的假设——观点会被高度主观和语境化。

今天，可以根据互联网上的信息创建新的指标，特别是在社交网络上。从社交网络中提取的信息创建新的指标，可以是检测人们生活和各种组织生活中重要事件的关键。通过组织中的企业事件，可以根据事件类型和事件中隐含的情绪类型直接检测股票市场未来的波动。在下一节中，我们将通过社交网络推特展示事件检测领域的一些工作。

2.4 事件检测

这些年来，社交网络的活动率有了很大的提高。数以亿计的用户在推特等社交网络上注册。用户交换想法，并在推特上分享每天的信息。经常在世界各地实时描述。在事件监测和对发生的事件的感觉中，推文的增加已被证明是非常有用的。因此，当目标是检测社交网络并发现用户是否有积极或消极的信息时，情绪是必不可少的工具。

利用网络推特对自然灾害和社会事件的检测进行了广泛的分析和讨论。这些事件通常有几个属性1)是大规模的，有许多用户对事件感兴趣，2）由于各种原因影响人们的日常生活，因此他们被诱导发布关于事件的推文。这些活动包括社会活动，如大型聚会、体育赛事、展览、促销产品事故和政治活动。它们还包括自然事件，如风暴、大雨、龙卷风、台风、飓风和地震。某某某（2009）指出，当一个事件发生时引起恐慌，人们就会在社交网络上寻找信息。这篇文章引用了弗吉尼亚理工大学悲剧的例子，学生们可以在当局的前一天制定出完整的所有死去学生的名单。

正如我们已经知道的，社交网络已经成为现实世界事件的主要信息来源。大多数旨在从这些源提取事件信息的方法通常是用消息的时间背景。然而，利用地理信息系统的位置信息同样很重要，它可以探测到局部事件比如公共事件或紧急情况。用户发布与事件位置相近的消息像一个“人类传感器”一样描述事件。由于推特必须过滤出版的位置，许多研究都使用推特数据来源，以便通过位置获取大量数据从而能够探测特定城市或国家的事件。

在探索和研究从推特流事件检测的技术方面，已经有了几篇文章。这些技术旨在寻找在时空中展开的真实事件。这是一个巨大的挑战，在推特上检测事件能够处理大量信息而不含没有意义和充满噪音的内容。某等（2015）提出了一种根据事件类型、检测任务和方法对技术进行分类的新方法并讨论了常用的特征。

某等（2014）发表了一项研究，目的是研究推特在巴西公共政策发展过程中可能出现的登革热疫情检测工具的有效性。这项研究表明，推特显示出了登革热流行的检测工具。分析表明，与卫生部提供的数据相比，推特数据有相同的行为。在事件或紧急情况下，人们在社交网络上的行为也成为研究的主题。某等（2010）决定了如何通过对两个自然灾害的新闻转发来传播整个网络的信息，例如红河的洪水和俄克拉荷马州的火灾。推特上发布的信息也被用来预测一些地区发生的地震。某等在推特上通过检测关键字触发“地震”或“震动”等方法在推特上开发了识别地震事件的技术，这一事件必须事先知道而且应该用简单的查询来表示。

摘要文本利用情绪传播算法，将所使用的标记词之间的依赖关系进行了区分并将其区分为强化、修饰、否定、对抗和让步关系。这一研究显示出了比简单的方法更复杂和精确的方法。

相对于在普通字典上所做的各种研究并将其与在互联网上发现的单词进行数值极性的研究，这种方法比这些研究提供了更好的准确性。这种方法的准确性来自这样一个事实：这种方法没有哦使用已经定义好的字典，而是应用基于上下文的算法来自动创建每个特定上下文的字典。本研究的经验表明，新创建的词汇库在使用无监督方法时具有优越性。

1. 系统架构

该模型实现的主要指标包括用户在社交网络上发布的推文中隐含的感觉，它基于四个文本分析工具，这些工具可以返回日常表达的情绪。这种工具对文本的积极、中性、消极的情绪给出了一个相应得分。基本上，高分表明对公司积极态度的存在。系统架构如图一所示，由5个主要模块组成：金融社区、推特过滤器、情绪分析器、规范化和检测事件。收集的推文进入过滤器：只与特定的股票或指数（如道琼斯）相关的推文，下一步将推文组织成文件，我们的系统将用情绪分析器模块来分别计算使用每个情绪工具得出的每条推文的每日得分。在最后阶段，对文件执行标准化使推文的体积（volume）正常化。这种标准化是必须的因为随着时间推移推文的流入将越来越多。

系统的最后一个模块是事件检测，即一种与分析上面计算的分数有关的的方法，根据图的峰值来检测特殊事件的发生。高的峰值对应于高分数亦即用户满意度高的事件。反过来，较低的峰值对应负面分数，与引发金融界不满的事件的发生相应。

3.1推特

推特为开发人员提供了一个允许他们访问更新、状态、数据和用户的REST APIs（s.d）以及其他特征。REST(表示状态传输)是由各种基于网络的体系结构衍生而来的一种网络架构混合模式。它定义了一个连接器接口，允许客户“对话”到单向服务器。推特还通过APIs（s.d）提供了大量的实时信息。推文可以按#字符之前的标签文字分组，用于在推文中标记关键字或主题。此外，用户还可以在用户发布的推特上发布转发推文。推特是这个工作的数据挖掘源。APIs（s.d）可以提供对各种数据的访问。

1. 金融社区的定义

我们选择社交网络推特作为数据来源。我们用推特API作为实施一个金融社区的推文数据收集工作的软件。金融社区由推特用户的一个子集组成，他们在金融市场上有相似的兴趣。金融社区的理念是选择那些对股市感兴趣的用户并拥有更多的追随者。真正重要的不是这些人而是他们的影响力，我们对他们说什么感兴趣。

B)金融社区的建设

以10个很有名的代表投资专家、金融新闻提供者、管理人员和公司创始人的用户账户开始，最终的目标是找到跟随他们的人。用户账号选择的第二个阶段是基于这10个用户的追随者，他们很可能属于同一个金融投资界因此有相似的利益。图2显示金融共同体的选择。基本上，软件开发人员分别访问了10个有影响力的用户和他们的数千追随者。对每个追随者，算法分析了追随者的数量和追随者的追随者的数量。如果用户比他的追随者有更多的追随者，并且追随者超过100个用户，那么就把他加入金融圈，认为他是级别二的有影响力的用户。例如图二Ray Dalio是10个有影响力的用户之一，拥有6584个用户。我们的软件分别访问并分析了6584个追随者，追随对象比追随者多的用户被抛弃，不被纳入金融界。在筛选过后我们得到了由数千个id组成的金融社区，这些id是用于在特殊事件检测中使用的推文的起始点。

3.2过滤器

在这个阶段，我们为每一个金融社区的用户赋予一个推文文件。滤波结果如图三所示。有了数据库，就需要对所有内容进行过滤。第一个过滤器使用我们编写的字典，它包含与金融市场和股票交易相关的865个单词。在某些情况下我们需要为相关词汇添加些动词的共轭。这个字典把来自银行和投资领域的信息保存在一起，为用户提供了对所有最常见甚至不常见的金融术语的彻底可靠的含义。这个初始过滤器的目标是删除所有与金融市场无关的推文。基本上，传递到下一个筛选阶段的推文必须包含至少一个构成字典的单词（金融字典）。例如：资产、平均、商业购买、资本、投资、收益、现金、组织市场、崩溃、债务、股息、经济交流、通货膨胀、外包等。“买”这个词字典单词的词形变化动词“买，买的过去式，买的过去分词”也应该被包含进去。而如果一条推文包含单词“苹果”，但与该公司无关，这个词作为一个与金融内容无关的词就应该被抛弃。与“苹果”水果有关的推文并没有通过第一个过滤器。

然后另一个过滤器被应用于通过一个与道琼斯指数公司相关的关键词来过滤推文。这个过滤器基于一个文件，该文件包含每个公司的几个关键词。表2显示了30家公司的4个关键词。根据每条推文的关键词，过滤器在30个文件中组织推文，每个文件对应于道琼斯工业平均指数中的30家公司之一。在推特上，只需要在推文中找到关键词的文件中的一个关键词，就能与与其中一个公司匹配。该文件由三种类型的关键词组成：第一种是公司名，第二种标识为股票行情，它的特征是“$”的符号、它的股票代号。推特创造了“现金标签”作为人们创建关于股票市场的推文的一种方式。“现金标签”是“$”这个符号加上5个字母，是股票的象征。股票符号是分配给特定市场上每个证券交易的唯一标识符。例如，SAAPL是为苹果公司服务的。最后一种关键词是由四个符号然后加上公司名称组成的现金标签。

举个例子，推文“Fri Sep 06 21:12:21 BST 2013-Big news: $AAPL reportedly struck deal w/China Mobile to sell new cheaper iPhone to its 700 mil users”它被安放在Apple.txt文件中。

在这两个步骤的最后，我们得到30个文件，每个文件对应一个公司，它包含所有来自金融界的关于公司和金融市场有关的推文。

3.3 情绪的定义

MySentimentApi

表2是用四种不同工具评估6条推文的例子。我们分析推文中表达的情绪的应用是在Java中开发的，也是基于包含极性的字典中的单词。“情感140”是一个通过极性来对推文进行分类的Web应用程序。评价采用了远程监控方法（et al.2009）。“情感140词典 版本0.1”使用一个由160万条推文组成的样本，表情被当做积极或消极情绪的标签。