越来越智慧的 NLP舆情分析

近年来，自然语言处理（NLP）成为数据科学领域中最热门的主题之一,它被广泛应用于搜索、翻译、社交媒体监控、聊天、广告、招聘、语法检查等等，机器学习让结果更智能、更快速、更精确。如何将NLP结合机器深度学习应用于舆情（Public Opinion）分析，找出违约跟并购中的舆情因子，在金融市场为投资提供有价值的参考？《时代财智》与正在美国的新加坡国立大学商学院怡合集团金融讲席教授、亚洲数码金融研究所所长段锦泉教授相约越洋问答。

| 文：张军

舆情管理日益受到公司关注，而舆情监测、分析和决策方法显得错综复杂，那么机器是如何从自然语言进行舆情分析的呢?

首先，我们需要了解自然语言处理(NLP)。所谓NLP是指对人类语言进行如句子结构解析、词形标注、机器翻译和对话系统等的自动分析和表现的计算技术，NLP的整体研究方法包括语料的获取、预处理、特征化、模型训练和建模效果评估五个步骤，分析方法包括词法分析、句法分析、语用分析和语义分析，其中文本的语义分析是我带领团队的NLP研究重点方向。

使用NLP处理文本语义分析，首先要搞清楚目的，它是作为目标变量（因变量），还是一个解释变量（自变量）？这要求我们把文本这类非结构化数据，转换成一个类似商业

上的变量，成为一个结构性的数据（如解释变量），在进行金融分析（如预测）时，不能认为舆情是唯一的决定要素，舆情之外尚有好多其它的解释变量，舆情在金融分析里最多只是一个补强的功能，它无法取代传统方法的金融分析。

看看小孩子学习文法的过程，他们靠一个一个的句子，一段一段的文本不断地重复学习，反复中学会了句子在不同环境下使用的多样性；而这也正是现在自然语言处理方法的基本原理。通过不断地重复去读那些懂文法的人写得文本，当读了很多的文本后，自然而然，就从不懂文法开始学会了文法，机器正是通过同样的过程，了解了语义、学会了文法。

如果处理对象是声音，NLP如何进行处理？机器如何避免谐音所产生的误判？

NLP机器就是一个抽象的计算机，通过读大量的文本，赋予文字在句子里、句子在文本里的意义，让机器在特定、多样性的目标下进行学习，对文本里的内容按文本的内涵赋予相应的数字符号。

声音和文字的处理过程有所不同，对于声音档，可以先把它转成文本，然后再数字化，当然也可以跳过中间过程，直接从声音到数字化。这类转化不是100%准确，对于一些谐音也许会做出误判，就如同我们人工写文本，偶尔会写错别字一样，本质上没有太大差别。减少这样的错误，需要经过对特定目标的反复学习迭代。

建立舆情分析的模型逻辑是怎样的？偏差是如何修正的？机器经过一段时间的“学习”，是否会越变越聪明？

研究NLP的学习，分静态和动态文本方法。所谓静态文本方法，就是在文本收集过程中，不断地使用同样的资料，不断地去读它，这些资料是静态的，是不变的，而方法则在不断修正，因为技术改变，体会的深度和正确性也不断提升，输出的结果也不断改变，就像我们读《红楼梦》，读第二遍、第三遍的感受一定和第一次的体会不一样。所谓动态文本方法，是指方法固定，而内涵一直在变，文本一直在变。

在研究过程中，一方面，要在原有资料库上不断地改变方法，增加准确度。另一方面，从使用维度上，当输入为新文本时，输出也是新的。机器在更新，使用文本也在更新，同时也使用这个机器，进行再生产。

机器处理语言的能力是一个很漫长的开发、学习过程，我的团队目前还在研发阶段，研究结果没有真正上线，我们对应的是一个动态的文本。虽然背后的工作还没有对外公开，但是研究结果是公开的。团队现正在准备把舆情变成结构化数据加入运作系统，让每天发布的违约概率数据都包括这个变量。当然，这不是一个小工程。

NLP并不是挑战人类的机器人。推土机的产生是为了提高铲子的效率，至于修建公路用哪个，是由人而非机器来决定。

你们研究项目叫什么名称？是CRI项目吗？是否有时间表？

这个研究项目叫自然语言处理的另类资料（NLP’s Alternative Data）。我们的NLP团队所生产的“另类资料”将用于CRI（信用研究行动计划）平台，用来补强原本系统包含的结构性金融数据，所以这两个团队的最终目标是一致的。

目前，我们没有对外公布具体实践的时间表，因为现行的CRI线上平台，自2010年推出后已运行12年了，对全世界上市的公司的违约概率进行预测，每天都有新修订的资料出台，不断地技术改进也是CRI的常态。不过团队所从事的项目都有内部的时间表。我们将在一两年内正式推出、加入系统。但那并不意味着团队NLP研究、执行工作的结束，如同百度类的系统一样，内容不断在更新，方法一直在改进，资料也会越来越丰富。

您的分析研究，如何选用媒体平台做研究对象？如果媒体背后有利益集团所支持，会不会影响到预测企业违约的分析结果？

社交媒体里有海量信息，噪音也很高，甚至有故意操作或炒作行为；传统媒体具公众性，可信度也相对高，所以我们选择输入传统媒体的文本。因为媒体在新闻价值上的选择性，所以，这种媒体资料反映出来的舆情，只能够用来补强，不能够取代传统金融数据。

NLP工具被开发出来，能处理大量的文本，但并不是挑战人类的机器人。就像推土机，是为了提高铲子处理的工作量，至于修建公路是用推土机还是铲子，不是由机器决定而是由人来定夺的。

当自动化的程度越来越高时，机器能够做的事也相应变多。现在人类面临最大的问题是认知上的困难，以前机器做不到的事现在做得到了，人类感觉受到挑战。和传统研究一样，我们团队的研究经费主要来自公共部门，如国家基金委（NRF），新加坡金融管理局（MAS）和大学。目的是科学研究和为公共服务，研究成果也是公开的，并非为某一个利益集团而特别定制。

在舆情分析中，如果媒体内容变了，输出的内容就变了；我们并不能主导媒体产生的内容，而是翻译和诠释媒体所表达的意见。我们所做的工作并没有价值判断的问题，唯一的价值判断是是否采用某个媒体。

舆情分析是通过机器从媒体那里获取结构性数据的，现在已经出现机器人产生的数据文本，未来会不会变成了机器与机器之间的对决？

我不这样认为。当机器可自我学习、判断的东西时，行为是不可预测的，但随意的外插并不属于创造性的工作；苏东坡喝酒后能作诗，但这并不代表我喝酒后也能作诗。人类和机器的不同在于人类的创造性，机器可以做的事有规则性。

未来发展的方向是智能的自动化，制造出比较聪明的机器。自动化本来是很机械性的，发展到数字时代就开始变得比较聪明。当我们比较两台机器哪个更厉害时，不要忘了，其实是它们后面的团队哪个更厉害。

媒体是有立场的，机器如何知道？

当考虑到媒体的立场，其实是改变了目的，用NLP做舆情分析时，一定首先要搞清目的，工具的正确使用取决于目的。我带领团队的舆情分析，目的是补强企业信用分析的品质，手段是减少人的工作量。让人去读上百万篇的文章是不可能的任务，但机器不会取代人的判断。

不同媒体的文本所表述的意见可能不一样，本来就是正常的现象，NLP的技术相当程度上能够了解，不同作者到底想表述的是什么，关键是建模时不应加入主观的选择。主观选择往往产生很严重的偏差问题。建模后，让机器进行判断，不然就会产生主观选择。

舆情分析和公司违约风险的研究在全球发展的状态？请您介绍一下亚洲数码金融研究所目前处在一个什么水平？

亚洲数码金融研究所的CRI团队，做全球上市公司的信用分析，建有全世界最大、先进的平台。许多金融机构使用我们的方法和产品，比如国际货币基金组织，新加坡的金管局2021年的金融稳定报告，还有很多银行和保险公司也在使用CRI的产品和方法。

目前，CRI平台所产生的信用分析产品，完全没有用到舆情分析，也就是说舆情分析的结果目前还没有放入这个平台。由于传统的金融分析数据里不一定展现出舆情分析的内涵，所以我们NLP团队的舆情分析是准备补强用的。

我们的研究结果已经显示，舆情是有补强的作用，比如说2008年雷曼兄弟的案子，金融数字已经显现很糟的结果，而加入舆情分析的预测，则显示情况比金融数字的预测还要更糟，这一点已经得到了统计上的验证。

一个团队最糟糕的事情，就是大家想法是一模一样。我带领的CRI和NLP团队共约四十多人，融合了计算机、数学、统计、金融和经济等不同领域的人才，10年磨一剑，齐心合力解决金融问题，经过十几年的系统化努力研究，我们已建立了全球最大、先进的平台。时

关于段锦泉教授

段锦泉为新加坡国立大学商学院的怡合集团金融讲席教授，并担任国大亚洲数码金融研究所所长。段教授是威斯康辛大学麦迪逊校区的金融学博士，也是中央研究院院士、金融计量学会会士、国际信用资产经理协会的顾问委员。加入国大前，他任教多伦多大学罗特曼管理学院，为Manulife金融讲席教授。段教授于2009年开始推动信用研究行动计划，以公共产品的精神，提供全球134经济体8万余家上市公司，每日更新、机器产生的企业违约机率。段教授于2017年共同创办并担任，金融科技公司CriAT (创信) 的非执行主席，提供深度信用分析的服务。