自然语言处理技术在金融领域的应用

一、摘要

近年来，自然语言处理技术已经取得了长足进步，成为应用范围最广泛，也是最为成熟的 AI 技术之一。但实际上，自然语言处理技术在商业化应用上却鲜有突破性进展，真正成功或者能够完美满足人们日常生活需求的产品并不多。

而金融领域的NLP目前仍处于探索阶段，金融本身是一个专业性很高的领域，很多词汇在金融语境下会产生特殊含义，所有的子问题都会有一个独特的理解方式，而且金融领域衡量处理结果的方式也与其他领域不同。比如针对舆情分析，金融领域要求对市场未来的走势有一定的预见性。

二、自然语言处理的发展经历了哪些阶段？遇到了哪些挑战？

NLP的发展进程与人工智能发展的脚步大体相同，都经历了如下的发展阶段：

- 20世纪50 - 80 年代：简单的实现人类掌握的规则，基于人类的经验；
- 20世纪90年代 - 2000年左右：主要基于统计学的原理与方法；
- 2000年之后至今，由于数据的大幅增强、计算力的大幅提升，人们也逐渐开始将如日中天的深度学习方法引入到NLP领域中，在机器翻译、问答系统、自动摘要等方向取得了重大突破。

但同时也应当注意到，NLP目前也仍然面临诸多的挑战。人类的语言非常简练，在很多对话中是省略背景知识的。人类自己是可以很容易地理解这种省略的背景知识，但在NLP的过程中却可能是很大的挑战。

比如“司机，我在前门下车”这句话，当机器不了解具体语境的时候，就难以分清究竟在公交车前门，还是在北京前门站下车。

三、面向中文与英文的NLP存在哪些不同？中文NLP，特别是在金融领域存在哪些难点，有没有某种算法是最佳的？

从语言本身上来看，英文比中文更直接，利用名词就可以很大程度上判断出一句话的语义。作为表音文字，英文还可以通过语法、时态、词性、词根、词缀、单复数等形式来让机器判断真实意图。

中文是象形文字，没有各种词性的转换，也无法对某个单字进行拆分，因此机器一定要通过上下文语境来判断具体语义。由于中文的特殊性，同一个任务、同一个模型在英文语境的表现一般要比中文好。

中文分词是中文NLP的难点之一。如“结婚的和尚未结婚的”，应该分词为“结婚/的/和/尚未/结婚/的”，还是“结婚/的/和尚/未/结婚/的”，不同的分词方法会产生一定的歧义。再比如，“美国会通过对台售武法案”，我们既可以切分为“美国/会/通过对台售武法案”，又可以切分成“美/国会/通过对台售武法案”。

随着深度学习的普遍使用，中文与英文在语言上的差异也逐渐变成训练数据量上的差异，以往在NLP领域，可供使用的中文数据量比英文数据要少的多，这是目前中文NLP的难点之一。但是随着有越来越多的人投入到中文人工智能以及NLP领域的研究中来，中文数据集不足的问题正在逐年改善。

在金融领域，针对基础性问题，中英文所处的阶段其实大体相同，但是针对如情感分析、市场预测等复杂问题，由于要结合具体的语境以及相应的应用场景，同时要考虑训练的数量级问题，无论是中文还是英文的NLP要走的路都还有很多。

四、一个强大的NLP系统能够帮助金融机构解决哪些实际问题？

全网舆情监控、产业链分析、让机器帮助金融机构阅读大量新闻。

例如，商业银行希望使用更全面的数据进行企业的信贷风险管理，提前感知企业的潜在风险。目前常规的风险评估方法是根据企业公布的年报，并综合信贷员实地调查的结果进行判断，但是由于企业自身风险报出通常具有滞后性，公开信息覆盖度不高，看到的往往只是冰山一角，因此判断风险的手段十分单一。这也是NLP与人工智能可以发挥作用的地方。

NLP可以对信息进行多维关系的挖掘，评估企业之间的关系，并通过知识图谱直观呈现企业之间的关联，提前设立预警信号，一旦企业关系网内的相关对象出现任意变动，便可根据关系权重，快速地评估对整个关系网的影响程度。

根据上市公司公开财报进行产业链挖掘是对NLP的又一应用。产业链数据以所有A股上市公司财报为原始数据源，根据公开财报中的主营业务构成，提取关键词后输入至预训练的神经网络中，对其进行向量表达。接下来，我们对输入向量进行基于密度的聚类计算，输出不同密度的集群，并最终进行集群命名。

五、NLP 的优势：

目前复杂的分析程序特别擅长发掘电子表格和关系型数据中的重要关系，但是从可用数据来说，这些数据只占到了 20% 的比例。

商业的真正问题是如何从非结构化数据中提取有效信息 —— 比如浩如烟海的社交媒体文章、图片、邮件、文本消息、音频文件、Word 文档、PDF 文件以及其他信息来源。这些信息来源占到了另外的 80%，但是却不能被传统的计算机工具和方法来理解。

而且，相对于填写客户满意度调查，人们更加愿意在社交媒体上分享他们的经验。所以 NLP 在理解客户的真实感觉上将是公司的无价之宝。

六、当前的实际应用：

1、信用评分

信用评分是金融领域应用AI的最普遍的一个应用，它是一个用来识别与特定服务的申请人相关联的风险级别的决策支持工具，广泛应用于银行、保险公司以及其他希望识别客户质量与风险的金融机构【6】。传统的方法是使用支持向量机、神经网络、对数回归、通用编程、最近邻以及其他混合方法。通过引入 NLP 方法，可以提高信用评分精度。

2、情感分析

人们非常擅长于揣测情感。情感似乎是把声音语调、谴词造句以及书写风格组合在一起的东西。要让计算机真正理解人类日常交流的方式，它们不能仅仅停留在理解只言片语的字面含义上；它们还需要理解情感，需要理解我们真正的意思。

分别文本情感的任务非常有挑战。一段话经常是微妙的，比如一段负面的评价可能并没有大量的负面词汇出现，反过来对正面的评价也同样成立。研究表明通过使用 NLP，可以在使用支持向量机方法的基础上，再进一步提高分类精度。

3、客户服务

NLP 以及其智能语音助理现在已经是我们生活的一部分了，在呼叫中心，劳力成本将持续下降，人们会从重复劳动中解放出来，因为机器会让呼叫中心更加个性化、自动化、更高效以及更方便。最近比较大的进展是，除了呼入，还有呼出都已经是自动化的了。

对于客户服务，NLP 的应用大致在3个方面：

- 实时语音识别
- 对话管理
- 对话分析

4、数据分析与预测

大量的且不断增长的金融报告、发布会以及新闻，让公司希望能够自动分析并且保持业务竞争优势。

七、参考文献

【1】2018 年 NLP 应用和商业化调查报告. https://www.infoq.cn/article/BvlY5qB_pcZRMwsIRgGk

【2】金融领域中的自然语言处理. https://zhuanlan.zhihu.com/p/65313931

【3】ONLINE BANKING IS EVOLVING WITH NLP & DEEP LEARNING. https://medium.com/@teamrework/online-banking-is-evolving-with-nlp-deep-learning-24822f477289

【4】Natural Language Processing Applications in Finance – 3 Current Applications. https://emerj.com/ai-sector-overviews/natural-language-processing-applications-in-finance-3-current-applications/

【5】Natural Language Processing in Banking – Current Applications. https://emerj.com/ai-sector-overviews/natural-language-processing-banking-current-applications/

【6】Omar Ghailan, Hoda M.O. Mokhtar, Osman Hegazy． Improving Credit Scorecard Modeling Through Applying Text Analysis[D]． (IJACSA) International Journal of Advanced Computer Science and Applications，2016． https://pdfs.semanticscholar.org/965f/5fcc7695461839f5e1eeaf9c4feb02305f85.pdf

【7】Leandro Alvim, Paula Vilela, Eduardo Motta, Ruy Luiz Milidiú . Sentiment of Financial News: A Natural Language Processing Approach[D] . Brazil : Pontifícia Universidade Católica do Rio de Janeiro. https://pdfs.semanticscholar.org/e71d/dfebe61fd1d1e8ce9b2cb80a2e8c796a6e04.pdf