课题是微博谣言分析检测。

【下一页 2】

主要分四个部分展开。

【下一页 3】

首先是选题背景。

【下一页 4】

2018年下半年开始，企业生存越来越艰难，市场人才竞争越来越激烈，甚至有小道消息称腾讯进行了一批大规模裁员，人数高达6000，不少人开始警惕互联网寒冬真的来了吗？不过不久腾讯便进行了紧急公关，“腾讯总共才3万人，一下子就干掉两成，还让不让我们干活了？真相是：HR今年有3000多人的社招任务，忙得很，没空裁员。欢迎各位有志青年加入鹅厂。”

【下一页 5】

网络上未被证实或故意扭曲的言论会对个人、企业、社会造成极大的影响，谣言就是指这样的没有相应事实基础却被捏造出来并通过一定手段推动传播的言论。网络空间中的谣言大致有四类，一是**恶意营销类谣言**，商家借此事件热度来推销自己的产品，提高品牌知名度。二是**社会安全类谣言**，例如下雨之后的 “自来水污”、 “政治黑幕”等，引起社会恐慌；三是**生活常识类谣言**，例如食品安全、养生保健等，瞄准的是网民生活中的痛点；四是**爱心慈善类谣言**，例如 “无钱上学求助救济”之类，主要是利用网民的同情心。

【下一页 6】

随着各大社交平台的飞速发展，个人的表达更为自由，以微博为例，近年来微博的用户规模逐年增长，为谣言传播提供了良好的温床。一些用户有时从自身利害关系出发，演绎出虚假信息，发布到网络上，成为谣言源头。一些用户因为自己的不安全感和宣泄的需求，并未识别就直接转发，使得谣言在短时间内迅速扩散，造成巨大消极影响。各大社交平台也在抓紧防控谣言的传播，Facebook更是雇用数万人来进行人工审查，其中耗费的人力成本可想而知。

【下一页 7】

如何能够智能化的对消息的可靠性进行评估、缩小人工审查的范围？我们便想通过对微博数据进行分析，结合微博文本、用户、传播方面的特征，实现谣言的自动化识别或早期检测。

【下一页 8】

整体思路是首先仅仅对微博文本使用词袋模型进行分类；然后是构造基于内容、用户、传播的特征，建立逻辑回归和决策树模型进行分类；最后是考虑时间效应，将事件传播分成多个时间段，再进行建模。

【下一页 9】

接下来，我们将对此次使用的数据进行描述。

【下一页 10】

此次使用的数据集是由香港中文大学自然语言处理课题组在2016年发布的，收集了2012年2月至2015年12月期间的380万条微博，涉及274万个用户。谣言数据来自微博社区管理中心中不实信息一栏。非谣言的数据来自同时期一些较权威的微博号发布的微博，并进行了人工审核，并非不实信息。

【下一页 11】

下面我们对谣言或非谣言事件进行定义。事件是指一条原始微博的所引发的一系列讨论，包含了原始微博以及其被转发过程中所涉及的微博。如蒙牛“良心奶”事件（有网友声称蒙牛的生产日期竟印成了2月30日，带来了一系列社会恐慌），其原始微博为谣言事件的源头，其后包含103682条转发，用户常常在转发时附带自己的看法，如有人开始对大陆乳品缺乏信心，有人则觉得这张一年前的照片并不符合事实。对每个事件所包含微博的条数绘制了分布直方图，呈左偏分布，平均每个事件包含814条微博，大多事件包含讨论都不多，但也不乏如蒙牛“良心奶”一般的火爆事件。

【下一页 12】

接下来，我们进行了一系列的建模分析

【下一页 13】

谣言常常具有特定的文本特征，比如经常有人在网络上黑城管暴力制法，说“城管打人”了；有人总爱散播小道消息，“据爆料”，鸟叔上春晚1分钟要10万元；与人总是能找到“真相”，高考阅卷有“内幕”。

【下一页 14】

于是我们对4664条原始微博进行分词，对词频进行统计，并绘制高频词词云图。在谣言微博中，“城管”、“政府”、“高考”、“医院”、“地震”等词频频出现，这些词往往与政治事件、社会治安更加相关。而非谣言微博当中，“生活”、“喜欢”、“希望”、“发布”等词出现较多，这些词更加贴近人们的日常生活，随手传递正能量。

【下一页 15】

对于分词后产生的31973个词语，使用词袋模型表示文本特征，并使用朴素贝叶斯分类器进行分类预测。词袋模型即仅以某个词语出现的词数来表示这个词语的特征。朴素贝叶斯分类器即是，在给定词语的条件下，取使得这条微博存在概率最大的一个类别。

基于词频的朴素贝叶斯模型在测试集上的准确率是88.0 %。同时我们计算了所有词语在谣言中出现概率和在非谣言中出现概率之比，罗列出了比值最大和比值最小的五个词。更易在谣言中出现的“抽烟”、“袁裕来”、“刀”、“天然”、“揭露”这些词语往往和生活健康、道路安全、揭露真相、谣言传播者[袁裕来]更相关。更易在非谣言中出现的“①”、“哪些”、“住”、“小伙伴”、“收藏”等词语指向性不特别明显，但常在一些具有知识点[①②③]，安利内容推荐大家[小伙伴]来[收藏]的微博中出现。

【下一页 16】

但微博只有文本信息吗？

【下一页 17】

我们还可以考虑微博的用户以及微博的传播影响力。在文本信息提取时，我们还可以提取更加细粒度的特征。

【下一页 18】

在文本信息抽取时，就不再只考虑所有词语出现的次数。而是设定特征类别对文本类型以及情感指向进行提取。对于文本类型，可以提取文本的长度以及是否出现特定标志（微博中@是链接到其他用户，#代表参与的话题讨论，？代表质疑，！代表吃惊）。对于文本的情感指向，可以引入外部词典，统计文本中出现的褒义词、贬义词数量，还可以计算文本的积极、消极指数以及情感得分等等。

【下一页 19】

于是我们提取了微博中若干基于内容、用户、传播影响力的典型特征，绘制箱线图，这些特征在非谣言和谣言当中的值分布均有明显差异。

【下一页 20】

最终我们共构造了46维特征，其中包含28个单一特征以及文本在18个主题上的LDA主题分布。分别使用逻辑回归和决策树分类器进行建模。结果发现使用全部特征比单种特征效果更好，决策树模型在测试集上的准确率更高。

【下一页 21】

另外，我们还将逻辑回归模型中对每一维特征的逻辑回归系数提取出来分析结果。逻辑回归系数为正代表该特征的值越大越易预测为谣言，逻辑回归系数为负代表该特征的值越大越易预测为非谣言。挑选出逻辑回归系数值最小的前5个变量，可知微博当中第一人称数越多、携带链接、携带标签，微博用户通过认证并提供个人图片的情况更易在非谣言中出现。挑选出逻辑回归系数值最大的前5个变两个，可知微博中使用过多褒义词或贬义词、多问号或多感叹号的情况更易在谣言中出现。

【下一页 22】

不过谣言事件的发现通常需要人群的智慧，不实的信息在传播过程中往往会有漏洞流出，在判断一条消息是否为谣言时，还可同时考量其涉及的讨论。

【下一页 23】

我们统计了谣言和非谣言在传播阶段每一个时间段下，含“?”的微博的占总微博的百分比的变化，“?”常常代表用户对于微博信息的质疑，在对虚假信息的讨论中普遍存在。可以看到，在非谣言传播的过程中，“?”的出现频率并无巨大波动，民众在其中的质疑声逐渐衰落。在谣言传播的过程中，“?”的出现频率波动较大，谣言事件往往存在着事件发酵、冷静和探讨、转折以及多次发酵的过程。

【下一页 24】

在时间序列特征构造的过程中，我们对每一个事件，根据最后一条转发微博与原始微博的发布时间差，将传播过程划分为N个阶段，提取各个阶段中涉及微博的文本、用户、传播特征的平均值，并且计算相邻阶段相应特征之差并除以时间间隔作为新特征，将以上特征拼接起来，从而建模时间序列中的前后变化。接着，使用逻辑回归和决策树分类器进行建模。结果发现，使用全部特征的效果依然比仅仅使用单种特征的效果好，但由于特征维度过高，模型的泛化能力并不理想。

【下一页 25】

有趣的是，我们将特征在每个传播阶段的逻辑回归系数提取出来，发现了某些特征在不同传播阶段的影响方式。用户发微博数量这个特征往往反映了用户的活跃程度，在传播早期，逻辑回归系数为负，说明发微博数量这个特征值越小，越易被预测为谣言，也可以理解为就是在谣言传播早期更多的是发微博数量少的人（活跃程度低的人），此时可能是水军在散播谣言。接着，逻辑回归系数逐渐增大，说明活跃用户起到了推波助澜的作用。在传播后期，逻辑回归系数再次走高，说明在谣言传播中更多活跃程度高的用户参与进来了，此时可能有用户加入讨论进行辟谣。

【下一页 26】

传统分类器对于高维特征的泛化能力较弱，此时我们可以考虑适合于处理序列化数据的循环神经网络（RNN, Recurrent Neural Networks）模型。RNN可以看作是对同一神经网络进行多次复制，每个神经网络模块会把信息传递给下一个。由于基础的RNN的神经网络模块较为简单，在数据传递的过程中，容易产生梯度爆炸或梯度消失的问题，因此可以使用更为复杂的RNN单元，LSTM (Long Short-term Memory)以及GRU (Gate Recurrent Unit)。其中LSTM中有三个门函数，输入门、遗忘门和输出门，用来控制输入值、记忆值和输出值；GRU结构稍简单一些，含有两个门函数，更新门和重置门，来控制记忆值和输出值。使用循环神经网络的优势在于，一是使用词向量来表示文本引入了语义信息，二是能较好的保留长期记忆，三是避免了繁琐的提取特征以及手动提取特征时存在的认知偏差和不全面的问题。

【下一页 27】

在使用循环神经网络进行建模时，我们对每一个事件，根据最后一条转发微博与原始微博的发布时间差，将传播过程分为N个阶段，接着在每一阶段选取TF-IDF值最高的前k个词语映射为词向量后作为输入，最后输入神经网络模型中得到输出结果。若使用事件所涉及的全部微博，LSTM模型取得了最好的效果。

【下一页 28】

此外，使用基于时序特征的模型，可在传播早期诊断异常，从而实现谣言的实时监控与预测。对于测试数据，设定检测时间限，仅仅取距离原始发布时间在时间限以内的微博作为测试集，使用训练好的模型进行测试，观察模型在传播早期预测的准确率。可以看到，GRU模型能在传播早期（距离原始微博发布4小时内）达到较高的预测准确率，LSTM模型的预测准率稳步上升并逐渐赶超GRU模型。

【下一页 29】

【下一页 30】

最后，回顾整个建模过程，我们首先利用词袋模型提取出文本特征，建立朴素贝叶斯分类器，预测准确率为88.0 %。接着，我们基于内容、用户、传播影响力对微博特征进行了更细粒度的提取，使用决策树模型进行分类，预测准确率为89.2 %。最后，我们考虑时间效应，将事件传播分成多个时间段进行建模，在使用LSTM模型时，预测准确率达到了92.0 %。

【下一页 31】

但本研究的各个模型相对独立，未来可以考虑模型融合方法。在考虑传播效应的同时，还可以考虑到局部交互效应，将用户表达的支持、反对、质疑等态度考虑进来。目前还有很大一部分谣言属于文本与图片不符的类型，这需要提取微博中存在的图片的特征从而进行文本匹配。此外，随着知识图谱建立的更加完善，未来还可引入知识图谱对事件的客观程度进行评价。