

从标题文本中自动提取新闻价值

Alicja Piotrkowicz

利兹大学计算机学
院

scap@leeds.ac.uk

Vania Dimitrova 利

兹大学计算机学院

V.G.Dimitrova@leeds.ac.uk

卡蒂亚-马克特

海德堡大学计算机语言学研
究所(Institut für
Computerlinguistik)

markert@cl.uni-heidelberg.de

摘要

标题在吸引受众对在线人工制品（如新闻文章、视频、博客）的注意方面发挥着关键作用。对标题进行自动、大规模分析的能力对于促进大量数字内容的选择和优先排序至关重要。在新闻研究中，新闻内容已被广泛研究，使用人工注释的新闻价值--在对新闻项目的选择和优先级进行决策时隐含和明确使用的因素。本文首次提出了从标题文本中完全自动提取新闻价值的尝试。新闻价值的提取方法被应用于从《卫报》收集的大型标题语料库，并通过与人工标注的黄金标准进行比较来评估。一项众包调查表明，新闻价值会影响人们点击标题的决定，这支持了对新闻价值自动检测的需求。

，数字读者花了更多的时间来浏览、扫描和发现关键词。由Chartbeat进行的各种研究发现，38%的用户离开了

1 简介

在这个数字时代，"无限的媒体和有限的注意力之间的差距越来越大，使得任何东西都要吸引观众"（Webster，2014），头条新闻扮演着一个特殊的角色。它们的主要功能是吸引注意力，并作为在线数字内容的视觉入口（Leckner，2012）。这一点在社交媒体上得到加强，在间接参与的情况下（例如转发的新闻文章），标题往往是主要内容中唯一可见的部分。Liu（2005）发现，与印刷媒体相比

¹，而平均每个读者在一个网站上只花15秒钟²。美国新闻研究所的一项研究发现，大约有十分之六的人承认他们是“看标题的人”，只看标题而不看全文³。

因此，需要对标题进行自动处理，以促进对大量数字内容的选择和优先排序。这一点在新闻领域已经通过考虑新闻价值得到了研究。这些是一个事件的各个方面，决定了它是否被报道以及报道的程度，因此指导了编辑的选择。最近的新闻学研究（O'Neill and Harcup, 2009, p.171）表明，新闻价值也可以应用于受众接受的角度，从而帮助分析是什么吸引了受众对某些标题的关注。

从标题中自动提取新闻价值可以成为

一系列应用的核心工具。自动提取的新闻价值分数可以与在线关注度指标（如页面浏览量）相关联，以调查哪些标题方面会影响在线人气。它们可以在基于内容的推荐系统中发挥关键作用，特别是在没有用户模型的情况下（所谓的“冷启动”问题）。头条新闻价值的洞察力可以被纳入在线内容发布，例如YouTube⁴，以指导作者如何编写标题文本来吸引受众的注意力。此外，数字人文研究者可以对不同数字媒体类型、流派、人口统计学等的新闻价值进行大规模比较。

尽管新闻标题的价值很重要，但没有自动的计算手段来从标题文本中提取它们。这需要先进的文本处理来计算适当的

¹²

<http://slate.me/1cJ7b5C><http://yhoo.it/2cEQMV>
C³ <http://bit.ly/21LwfS5>

⁴<https://www.youtube.com/>

可以与新闻价值相关的特征。这使得问题具有挑战性，因为新闻价值往往涉及隐性知识。没有可以用于自动文本处理的新闻价值的精确定义，这一点因标题文本的性质而变得更加严重。重要的是，没有研究表明如何将新闻价值与可以从标题文本中自动提取的各种特征相联系。

为了应对这些挑战，我们利用最先进的技术来开发一种从标题文本中自动提取新闻价值的方法。我们的解决方案包括几种NLP方法，如维基化、情感分析和语言建模。我们进一步将它们与其他人工智能方法相结合，包括一个突发检测算法，以提出估计实体突出性的新技术。该方法在一个著名的新闻来源--《卫报》的大型新闻头条语料库中得到了应用和评估。

本文以头条新闻的价值为重点，提出了一个处理数字内容的新视角，并通过文本分析作出了贡献：

(i) 提供了第一个从标题中完全自动提取新闻价值的计算方法，该方法结合了相关的NLP技术；(ii) 通过将计算方法应用于大型新闻标题语料库，并将自动注释与为该任务开发的黄金标准进行比较，评估了新闻价值特征的设计；(iii) 通过用户众包研究证实，人们选择点击新闻项目是受标题中的新闻价值影响的，表明自动新闻价值检测的重要性。

2 相关工作

在NLP界，标题作为一种文本类型被单独研究的地位越来越高。此前的研究表明，标题可以独立于全文发挥作用。根据Dor (2003)的研究，读者在阅读标题的时候会得到“最好的结果”。经验研究似乎支持这一点--Gabelkov等人 (2016) 发现，Twitter上59%的共享新闻内容没有被点击，也就是说，在被分享之前没有被阅读。这使得头条新闻成为社交媒体

上分享内容的关键。在期刊界，标题的重要性已经得到了认可。例如，Althaus等人 (2001年) 研究了全文的替代物。

包括标题在内的文本以及它们对结论分析的影响。Tenenboim和Cohen（2013）对标题内容对点击和评论的影响进行了研究。然而，这些努力包括手动注释，这限制了他们的范围。最近，NLP研究人员也关注标题，包括标题的生成（Gatti等人，2016）和用于流行化内容的关键词选择（Szymanski等人，2016）。我们通过提出新闻价值来分析标题，对这一正在进行的NLP研究进行补充。

新闻价值起源于Galtung和Ruge(1965)在新闻研究领域的工作。此后，人们提出了各种新闻价值观的分类法：Bell (1991), Harcup and O'Neill (2001), Johnson-Cartee (2005) and Bednarek and Caple (2012)。尽管在颗粒度和定义上存在差异，但所有这些分类法之间都有相当大的重叠。这使得我们可以选择最常被提及的、与标题文本最相关的新闻价值。这些价值包括：突出性、敏感性、超然性、接近性、惊奇性和独特性。我们提供了一种系统的、完全可重复的方法，从标题中自动提取这些新闻价值。此外，我们表明，这些新闻价值影响人们点击标题的决定。

新闻价值已被广泛用于新闻学研究，但研究人员仍主要依赖人工注释。例如，Bednarek和Caple（2014）使用新闻价值来分析新闻话语，而Kepplinger和Ehmig（2006）使用它们来预测新闻文章的新闻性。由于新闻价值需要手动注释，新闻研究中对新闻文章的大规模分析主要集中在通过人工元数据容易获得的方面（如Bastos（2014）中的主题）。在使用编译方法对文本进行大规模的新闻价值注释方面，已经有一些有限的尝试，但这些尝试最多只能说是半自动

的。例如，Potts等人（2015）从一个预处理过的语料库中手动选择新闻价值指标；此外，该方法依赖于关键词，并依赖于主题。本文首次尝试对新闻价值进行全自动的、与主题无关的提取，并在"大报"新闻来源的头条新闻上进行了验证。我们的新闻价值检测在很大程度上不是针对新闻的

并可扩展到其他类型的书目。

从NLP的角度来看，头条新闻构成了一个工程挑战。这包括语言方面，如时态的不寻常使用（Chovanec, 2014）和故意的模糊性（Broderick和Coulson, 2010）。还有一些特定领域的现象，如点击诱导（Blom和Hansen, 2015）。头条新闻通常很短，这限制了许多NLP工具所依赖的上下文数量。虽然对标题的特征工程研究较少，但有一些研究工作是专门针对短文的。推特已经引起了相当大的关注，导致了一些特定的Twitter工具的发展（例如TweetNLP⁵）。Tan等人（2014）是一个从推文中进行特征工程的例子，专门研究了措辞及其对流行度的影响。另一个与标题密切相关的文本的例子是在线标题，例如Reddit上的图片标题（Lakkaraju等人, 2013）。许多方法包括各种语音部分的比例、情感与语言模型的相似性等特征。然而，它们需要被调整以适用于头条新闻。例如，由于头条新闻提供的内容有限，在词级上进行的情感分析更合适（参考Tan等人（2014），Gatti等人（2016），Szymanski等人（2016））。对于每个新闻价值，我们要么重新实施最合适的最先进的方法，要么实施新的技术，以很好地处理标题。

3 新闻价值的提取

我们提出了六个新闻价值的特征工程方法。

之所以选择这六个，是因为

它们经常出现在新闻价值分类法中

（参照第2节）。表2总结了特征计算的方法。

尽管我们的目标是一个通用的框架，但我们受到了新闻领域研究的启发。因此，这些特征是由与新闻内容相关的新闻价值提供的。

预处理。 所有的头条新闻都被标记为语音部分（Stanford POS Tagger (Toutanova et al., 2003)）和解析（Stanford Parser (Klein and

符号化。 我们将标题 H 视为从POS标记器中获得的一组标记。我们把 H 中的内容词集合表示为 C ，把 H 中的实体集合表示为 E （参见表1）。

表1：预处理： H （标记集）， C （内容词集）， E （维基实体集）。

"艾玛-沃特森的化妆推文突出强调了修饰之美"

$H = \{ Emma, Watson, 's, makeup, tweets, highlight, the, 商品化, 的, 美 \}$

$C = \{ 化妆, 推特, 亮点, 商品化, 美丽 \}$

$E = \{ 艾玛-沃特森, 商品化 \}$

NV1：突出性。 对杰出人物（精英国家和人物（Galtung and Ruge, 1965），以及最近的名人（Harcup and O'Neill, 2001））的提及是关键的新闻价值之一。

我们将突出性近似为一个实体在网上得到的关注量。由于在线关注度随时间变化，我们考虑长期与近期的关注度和突发性。我们通过使用维基化来获得实体并考虑它们的爆发性来扩展之前的工作。

对于一个实体 e ，我们将其表示为 $pageviews_{e,d-m,d-n}$ ，即该实体的维基百科每日页面浏览量的中位数⁷

$d-m$ 和 $d-n$ 之间。日期编号是参照文章发表日 d 确定的。维基百科的长期突出性是经过一年的计算($pageviews_{e,d-365,d-1}$)，以及维基百科最近在当天的突出表现

出版前 ($浏览量_{e,d-1}, d-1$)。⁸ 对于一个从以新闻为中心的角度看突出性，我们还计算了 e 在新闻中的提及次数之和。

来源于出版前一周的头条新闻

日，表示为 $newsmentions_{e,d-7}, d-1$

Manning, 2003))。Wikification（一种将文本中

的关键词链接到相关维基百科页面的方法；例如

Mihalcea和Csomai (2007)) 被用来识别文本中的实

体。头条新闻是维基化的

由于实体表现出不同的突出时间模式，我们将实体区分为 *稳定的* 突出（如 SILICONE）和 *突发的* 实体，即在短时间内突然突出（如 EBOLA VIRUS）。为了识别突发的实体，我们采用了 Vlachos 等使用 TagMe API⁶，这个工具旨在为短途旅行提供服务。
文本，使其适用于标题。

⁵<http://www.cs.cmu.edu/ark/TweetNLP/>

⁶<http://tagme.di.unipi.it/>

人（2004）的突发检测算法（参见算法1）。如果一个实体在给定时间段内的移动平均数高于临界点，则被定义为 *处于突发状态*（参见图1）。我们以两种方式使用实体突发。首先，爆裂性表示 e 处于的天数。
一年内的爆发（*天爆发* $e, d-365, d-1$ ）。第二章

⁷<http://dumps.wikimedia.org/other/pagecounts-ez/>

⁸我们发现前一天的突出值最接近当天的实际突出值。

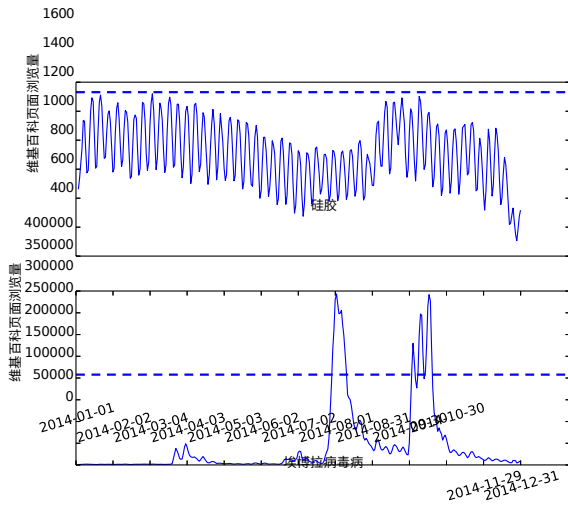


图1：两个实体的维基百科页面浏览量移动平均数（MA）的时间序列图：非爆发性的硅酮（顶部）和爆发性的EBOLA病毒病（底部）。虚线表示爆发的截止线。

另外，当前的突发大小表明任何 e 在发布前一天处于突发状态，比MA高出多少标准差 e （ $dayburst_{e,d-1,d-1}$ ，如果 e 处于突发状态则返回1，如果不在则返回0）。我们是第一个考虑到流行预测的突发性的人。

算法1 突发检测算法改编自Vlachos等人（2004）。经过实验，移动平均数被设定为3天，截止点为标准差的2倍。1: 计算实体 e 的长度为3的移动平均数 (MA_e)

为序列 $d-365, \dots, d-1$ 。

2: 设定截止点 = 平均值 $(MA_e) + 2 \times SD(MA_e)$ 3: 突发 = $di / MA_e(i) > \text{截止点}$

由于一个标题可以有多个实体，所有的突出度量都是通过对 H 中的所有实体进行求和来汇总的（见表2）。

NV2：情感。这指的是带有情绪的事件（Johnson-Cartee, 2005）和使用带有情绪的语言（Bednarek和Caple, 2012）。与情感和情绪有关的特征已被证明可以影响新闻文章的病毒性（Berger和Milkman, 2012）。然而，这种影响还没有针对头条新闻进行研究。

我们测量标题中具有正面或负面含义的词（使用内涵词典（Feng等人, 2013））。其次，我们测量有偏见的内容词的百分比（使用偏见词库（Re-casens等人, 2013））。例如，同一个政治组织可以被描述为**极右翼**、**民族主义**或**法西斯主义**，这些词中的每一个都表明对某种解读的偏见。

NV3：超水平性。事件的规模（Johnson-Cartee, 2005, p.128）或大小（Harcup and O'Neill, 2001）被认为会影响新闻选择。

我们专注于明确的语言学指标

作为情感的直接衡量标准，我们结合了SentiWordNet（Baccianella等人, 2010）的内容词的积极性和消极性得分，并按照Ku-cuktunc等人（2012）的方法计算晚期情感和极性分数。情感也可以是直接的。首先，一个词本身可能是客观的，但带有负面的内涵（如**尖叫**）。因此，我们衡量内容词的百分比

事件大小：比较级和超等级（用语篇标签表示），以及扩大级（用增强级和下降级表示）。对于后者，我们结合 Quirk 等人（1985）和 Biber（1991）的列表，得到 248 个增强词和 39 个减弱词的词表。

NV4：接近性。这种新闻价值被解释为事件在地理上（Johnson-Cartee, 2005, p.128）和文化上（Galtung and Ruge, 1965）与新闻来源或读者的接近程度（Caple and Bednarek, 2013）。

根据一个假设，来自新闻机构所在国家的读者构成了其读者群的主要部分，我们关注与新闻源的地理接近性。我们使用一个二进制特征来表明一个标题是否指的是一个在地理上接近新闻来源的实体，并手动创建一个包括国家、地区、首都（总共 17 个与英国有关的术语）名称的词表。然后，我们在标题文本（"伦敦烟雾警告，撒哈拉沙尘席卷英格兰南部"）或维基百科中提供的每个实体的分类（标题 "低估皇家邮政股份，一天内使纳税人损失 7.5 亿英镑" 的类别 POSTAL SYSTEM OF THE UNITED KINGDOM）中寻找匹配。

NV5：惊奇。涉及 "惊讶和/或对比"（Harcup and O'Neill, 2001）的事件会成为新闻。标题中的惊奇可以是隐性的（"丹佛邮报雇用乌比-戈德堡为大麻博客撰稿"），这需要世界知识来识别它，也可以是显性的（"养蜂人创造了活蜜蜂的外套"），它产生于不寻常的词汇组合。

我们以明确的惊喜为目标，通过参考大型语料库，计算标题中短语的常见性。我们首先提取以下类型的短语：subj-v, v-obj, adv-v, adj-v

表2：《卫报》上的特征实现和统计。符号见表1。措施：中位数和最大值，流行率（非零分的比例），以及比较人工黄金标准和自动提取的Kruskall-Wallis检验（* $p<0.05$ ，** $p<0.01$ ，*** $p<0.001$ ）。

特征名称	实施	中位数	最大	患病率	KW
实体数量	$ E $	1	8	79%	***
维基百科当前的突发大小	$\frac{daysburst_{e,d-1} \times pageviews(e,d-1) - mean(MAE)}{SD(MAE)}$	0	57.16	12%	0.2
维基百科爆裂性	$\sum_{e \in E} \frac{daysburst_{e,d-1}^{-3}}{e^{0.2}}$	21	156	78%	***
维基百科前一天的显著性	$\sum_{e \in E} \frac{pageviews_{e,d-1}^{-1}}{e^{0.2}}$	1,642	1,031,722	78%	***
新闻来源 最近的突出表现	$\sum_{e \in E} \frac{新闻报道_{e,d-1}}{e^{0.2}}$	0	122	50%	**
NV2的情绪	- 最大积极性 - 最大消极性 - 2	-2.1	100%	0.1	
极性	最大阳性+ 最大阴性 -	0.5	1.88	79%	**
内涵	含有积极或消极内涵的w内容词 $ C $	0.34	1	92%	0.2
偏见	#有偏见的内容词 $ C $	⁹ http://www.nlp.cs.nyu.edu/wikipedia_data			
视觉	比较/超限	¹⁰ 我们用其他语料库和指标进行了实验，发现维基百科和对			
效果	强化器	¹¹ 实体重叠有助于确保标题是同一故事情节的一部分；包括			
近距离	如果在H或维基百科分类标签中明确提到英国，则为1，否则为0	0	1	35%	***
NV5 惊喜	$minLLp$ ，其中 LLp 是H中一个短语的对数可能性。	4.15	2,726,186	100%	*
NV6 独特性	$max_{t \in [0,72hr]} 余弦相似度(H, pastH_t)$	0	0.83	13%	*

N，N-N；并且用它们的转折形式生成一个正则表达式（例如，男人喝酒→男人喝酒|喝|喝）。对于每个正则表达式，我们从维基百科语料库⁹，并将每个短语的计数相加，计算出其对数可能性（LL）。特征值被赋予标题中最低的LL（因为我们正在寻找最令人惊讶的短语）¹⁰。

NV6：独特性。新闻必须是新的--"任何新的评论或情况[.....]都是对辩论的补充"（Conley and Lambie, 2006）。对头条新闻语料库中几个故事情节分析表明，在两个非常相似的头条新闻中，后一个往往不太受欢迎（"渡轮灾难：渡轮灾难：韩国总理辞职"比后来的"韩国总理就渡轮沉没事件重新签字"更受欢迎）。

对于一个标题H，我们选择H发布前72小时内的过去的标题，这些标题至少有一个TagMe实体重叠或都没有任何实体¹¹。对于一对H和pastH向量（使用tf-idf加权的Gigaword语料库创建），我们计算其余弦相似度。最高的余弦相似度被指定为特征值。

4 应用和评估

我们将特征提取方法应用于英国主要报纸《卫报》的头条语料库。这提供了一个广泛覆盖各种主题和体裁的时代，允许对新闻价值进行良好的探索。新闻价值的自动提取与人工标注的黄金标准进行了比较。

标题语料库。头条新闻语料库是使用卫报内容API 建立的¹²。我们向下加载了2014 年4 月期间发布的所有头条新闻

这确保了更多的报道。收集以往的头条新闻72小时的效果比其他截止点更好。

，产生了一个由11,980 个头条新闻组成的语料库。

自动注解。每个标题的特征值都被计算出来。表2 重新列出了《卫报》语料库中前述特征的统计数据（中位数、最大值、普遍性）。

人工注释的黄金标准。对于每个新闻值，我们从头条新闻语料库中选择了20 条头条新闻。为了使用最清晰的例子进行更准确的注释，我们从最高四分位值中随机选择了10 条头条新闻，从最低四分位值中随机选择了10 条。对于分为多个特征的新闻价值（NV1: 突出性，NV2: 情感，NV3: 超然性），特征组向量被排序以获得四分位数。总的来说，总共选择了120 个标题进行人工标注。三位专家注释者，语言学的博士生，将每个标题注释为正面或负面。

¹²<http://www.theguardian.com/open-platform>

对于前五条新闻的价值来说，是否定的（Y/N）（参见表3）。对于NV6:独特性，注释者得到了语料库中的20个标题，以及过去20个标题独特性得分最高和最低的标题（随机抽样）。注释者指出是否有任何过去的头条新闻与某一特定的头条新闻非常相似（即高度相关）。

注释者之间的一致性。注释者之间的一致性是用Fleiss的Kappa计算的。其范围从NV1:突出性（0.76）和NV6:独特性（0.73）的实质性，到NV3:超然性（0.43）、NV5:惊奇性（0.48）和NV4:接近性（0.55）的中等，到NV2:情感（0.22）的一般。注释者说，有时他们“凭直觉”选择，他们的反应可能每天都不同。这突出了自动检测新闻价值的挑战，因为新闻价值在某种程度上是被默许的。注释者的判断是通过多数投票汇总的，从而形成黄金标准。

与黄金标准的比较。我们计算了每个特征和相关的人工标签（例如，特征数和突出性，偏见和情感）之间的配对比较。用Kruskal-Wallis检验来确定两个人工注释标签（Y/N）的特征值的差异是否显著（参见表2的KW列）。这些结果表明，为某一特征计算的值是否正确反映了人类专家产生的黄金标准中的新闻值的存在。下面将讨论评估的结果。

5 关于特征提取的讨论

我们使用了一个新闻语料库，该语料库在“大报”（相对于风格和语气不同的小报）的范围内代表了广泛的新闻出版物。*卫报*语料库是一个可免费使用的资源，有助于方法和研究结果的复制。虽然对特征提取的评估是在一个语料库中进行的，但我们也将这种方法应用于另一个公开的“大报”语料库--《*纽约时报*》（参见附录A）。我们将在下面讨论《*卫报*》的评估研究结果，并将参考《*纽约时*

报》的特征提取结果来说明两个语料库的特征行为。

NV1: 突出性是最重要的问题之一。

我们使用维基百科的方法证明是非常可靠的。它出现的频率很高--《卫报》语料库中的大多数头条新闻都至少有一个实体（实体的中位数）。

= 1），这吸引了相当多的在线关注（维基百科长期突出度中值=1,342页浏览量）。一些标题包括非常突出的实体（维基百科前一天突出度最大值=1,031,722）。《纽约时报》的结果也很相似--每个标题都与至少一个维基百科实体有关（实体数量的普及率为100%）；在66%的标题中，维基百科的突发性、长期性和前一天的突出性得分都不为零。这表明，维基百科为计算突出性提供了广泛的覆盖面。维基百科当前的突发性大小是一个罕见的特征（在《卫报》中为12%，在《纽约时报》中为10%），因为在突发性中捕获一个实体是不常见的，因为突发性并不适用于所有实体，也不经常发生。

突出性的IAA是最高的（ $\kappa=.76$ ），与人工注释相比，几乎所有的特征都达到 $P<0.001$ 。这有力地支持了我们对突出性的实施，特别是使用维基化和维基百科作为突出性的来源。爆发性提出了一种看待突出性的新方法。虽然突发性（即一个实体在一年内有多少次明显高于其平均水平的页面浏览量）是一个可靠的特征，但当前的突发性大小（即文章发表前一天的突发性大小）与黄金标准没有显著关联。

NV2：情感是最难实现的新闻价值之一，因为它不是大报的典型特征，而且标题中带有情感色彩的语言并不总是能准确反映真实的情感或情绪。大张旗鼓的报纸的标题往往是相当中性的（情绪中值=-2；极性中值=0.5）。《纽约时报

的情况也是如此（情感=-2；极性=0）。然而，大多数标题至少包含一个有内涵或有偏见的词（内涵的普遍性=92%，偏见的普遍性=61%；《纽约时报》略低：78%和51%）。

IAA是公平的， $\kappa=.22$ 。许多标题是中性的，这一事实可以解释低同意率的原因，因为中性案例是专家更容易产生分歧的地方。此外，虽然对情绪的一个方面，如积极性/消极性的人工注释可以达到很大的一致（在Snow等人的研究中，专家之间的一致为0.76。

表3：注释的标题的例子。Y/N：多数人投票的人工注释。下图：通过特征组的汇总自动提取的数值（特征值范围参见表2）。

#	标题	突出表现	感受	超越性	靠近	惊喜
E1	"对EE/Orange的客户服务真的很纠结"	Y 0	Y 3	Y 0.125	Y 0	Y 3.23
E2	"珠峰雪崩造成至少12名尼泊尔登山者死亡"	Y 13272	Y 4.25	Y 0.17	N 0	N 4.15
E3	"为外国专家欢呼。毕竟，他们比我们自己的好"	N 672	Y 2.75	Y 0.2	N 0	Y 398
E4	"牧师；马丁-艾米斯的英格兰；以及《非常英国的文艺复兴》：电视评论 - 视频"	Y 36236	N 2.45	N 0.08	Y 1	N 4.15
E5	"本周新的现场喜剧"	N 0	N 3.25	N 0	N 0	N 102

al.(2008))，我们对情感的定义更广泛。注释者指出了表达情感的一个有趣的特征。一方面，有一些高度诱人的标题描述了一些悲惨的新闻事件（+情感，+情绪）。另一方面，有些标题使用了带有感情色彩的语言，但没有达到同样的唤起程度（+感情，-感情）。例如，*喜剧*（表3中的E5）有积极的情感，但没有唤起积极的情绪。与人工注释相比，四个情感特征中的两个达到了显著水平，因此我们的实现确实捕捉到了情感的某些方面。从标题中提取情感是一个挑战，因为它们都是短文，上下文有限，而且情感往往是隐含的，或者需要世界知识来识别（例如“几内亚的埃博拉病毒爆发：病毒是什么，正在做什么？”）。将情感和情绪分开可能会描绘出更清晰的画面。

NV3：超纲性是罕见的，但也是可靠的。它是最不普遍的新闻价值（在4-10%之间；在纽约时报中在3-6%之间）。平均值也都是零。我们对事件的定义比较狭窄，这可能是原因，但是我们决定把重点放在事件规模的明确的语言指标上（例如，*非常*，*几乎没有*），以保持实施与主题无关，更容易推广。

IAA是中等的（κ=.43）。三个特征中的两个在 $P<0.001$ 时是显著的。这证实了我们依靠POS标签和词表的方法确实捕捉到了这种新闻价值。唯一没有达到显著水平的特征是downtoners。Downtoners是一类旨在削弱它们

所描述的词的词（如*近乎*、*勉强*、*只是*）。它们不仅罕见（发生率为4%），而且还需要特定的知识来识别它们（我们识别了39个降调词，而强化词有248个）。考虑到

如果用一个更全面的词表来增加其覆盖面，那么downtoners可能會有更大的影响，其他的Superlativeness特征（comparative/superlative和intensifiers）可以可靠地用于头条。

NV4：近似度不高，但我们使用词表和维基百科分类的方法被证明非常可靠。这种新闻价值出现在35%的头条新闻中。这并不奇怪，考虑到《卫报》的受众遍布全球，因此大多数新闻并不是针对英国的（《纽约时报》的预估值与此类似，为32%）。

IAA是中等的（ $\kappa=.55$ ）。该特征在 $p<0.001$ 时达到显著性，因此我们捕获Proximity的方法得到了很好的支持。使用实体类别可以确保覆盖面更广，而且比使用词表更省力。这又取决于NER/wikiification工具的可靠性。在某些情况下，一个实体可能会被遗漏（参见表3中的E1，其中EE/Orange被遗漏，因此突出度和接近度得分都是零）。值得注意的是，接近度包括地理和文化上的接近。我们的注释者是英国居民，熟悉《卫报》，但读者的人口统计学可能会影响他们对一些实体的熟悉程度。在我们未来的工作中，我们会加入一些人口统计学的数据来深化Proximity的实现。

NV5：由于标题文本的特殊性，惊喜是很难实现的，但我们的方法是使用基于维基百科的语言模型来捕捉惊喜的措辞。这个特征的中位数对数可能性很低（4.15；纽约时报为4.04），这意味着大多数标题都有相当令人惊讶的措辞。这可能是因为头条新闻并不倾向于严格遵循日常语言的惯例（例如经常使用非时态动词和“我”字）。

名词集群)。当使用一个不是专门用于标题的语料库时(我们使用维基百科),对数可能性将趋于降低。

IAA是适度的, $\kappa=0.48$, 并且该特征是显著的 ($P<0.05$)。这表明,使用基于计数的方法可以捕捉到这种新闻价值。在其他类型的文章中,惊喜可能起到很大的作用,这种方法可以通过使用特定于标题的语料库或建立考虑到句法结构的语言模型来扩展。

NV6: 独特性, 或者说缺乏独特性, 是相当罕见的, 但我们的实现可靠地识别了这种情况。其发生率相当低(15%; 但在《纽约时报》中略高, 为34%), 这符合新闻业的基本原则, 即新闻必须是新颖的。

IAA是相当可观的, $\kappa=0.73$, 而且该特征是显著的 ($P<0.05$), 所以我们可以确定任何类似的标题都被识别出来。对具有非零独特性值的标题进行分析后发现, 大多数的标题要么是一个常见特征的一部分(例如"评论综述"), 要么是关于同一事件的持续故事情节的一部分(通常以一些媒体如视频为特色)。

总的来说, 评估的结果是令人振奋的: 对于每个新闻价值来说, 大多数的特征都能显著地区别于人工注释的标签。这意味着我们的方法至少成功地识别和量化了每个新闻价值的某些方面。

该研究还指出了需要进一步调查的公开问题。首先, 研究结果高度揭示了分析头条新闻时世界知识的重要性。例如, 对于像情感分析这样成熟的NLP主题, 我们发现虽然纯语言学的方法可以捕捉到标题中的大多数现象, 但它们在识别实体(如埃博拉病毒)中的情感方面还不够。同样地, 一个更通用的接近度方法需要世界知识来检测一个实体与读者的位置有关。我们将在未来的工作中解决这个问题。其次, 探索如何将所提出的方

法应用于其他类型的新闻来源(如小报)和新闻以外的类型将是很有趣的。除了新闻来源的突出性和独特性之外, 我们的特征并不针对新闻。其他类型的数字内容(博客、视频)的标题也包括突出的内容、情感或强化语。新闻价值检测为其分析提供了一个新的视角。

第三，我们的方法可以适用于其他语言，只要有一定的NLP资源（POS标记器、NER、情感词典）。这将使大规模的标题分析沿着多个轴线进行，如语言和流派。

6 新闻价值观是否会影响人们对标题的选择？

为了显示自动新闻价值提取对一系列应用的重要性（参见第1节），我们研究了新闻价值对普通受众是否重要。这一点通过一项众包研究进行了探讨。

调查内容。调查由五个简短的部分组成，涉及新闻价值NV1至NV5（NV6：独特性不包括在内，因为我们决定将重点放在单一标题所表达的新闻价值上，而独特性特征需要比较标题）。在每个部分中，我们向参与者提供了一个简短的定义和几个例子。然后，他们被问及以下问题：“在点击标题时，我个人会考虑这个新闻价值”，并给出五个李克特量表的答案（参见图2）。我们收集了标准的人口统计信息（年龄、性别、居住国、母语、新闻阅读习惯）。

参与者。众包平台CrowdFlower被用来招募调查参与者，使我们能够在全全球范围内收集答复，从而反映出在线新闻机构受众的全球性质。调查大约需要10分钟来完成，参与者获得2美元的报酬。在收集到的100份答复中，有96份被记录为完整的。虽然回答的质量普遍很高，但我们还是进行了一些质量控制。我们删除了任何超过75%的答案是中性的答复，以及完成时间处于最低四分之一的答复（以确保参与者花时间去理解这些概念）。在采取了质量控制措施后，有71人被选中：48名参与者是34岁或

以下，23名是35岁或以上；17名是女性，54名是男性；30名是英语为母语的人，41名是英语非母语的人；44名参与者每天阅读新闻，27名每周。

结果和讨论。结果显示在图2中。新闻价值对调查参与者的总体影响被表示为非常积极。NV1:突出性，NV4:接近性、

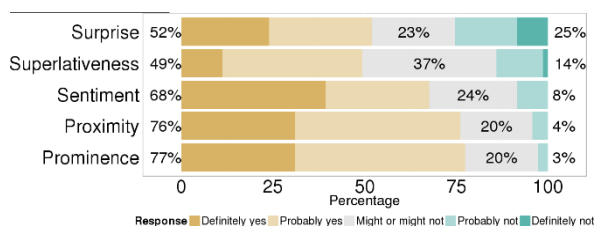


图2：对“我在点击头条新闻时通常会考虑这个新闻价值”这一问题的调查结果（N=71）。百分比显示的是正面、中性和负面的回答的总和。

和NV2:Sentiment的正面回答比例最高（分别为77%、76%和68%）。这遵循了新闻研究文献，其中这三种新闻价值可能是最重要的焦点。与黄金标准的比较证实，我们对NV1:突出性和NV4:接近性的实施重新反映了专家的判断。由于这次调查强调了情感的作用，我们有动力进一步发展它，以捕捉它的全部优点。NV3:超然性得到了最多的中性反应（37%）。一方面，这可能是由于这一新闻价值略微难以理解¹³。另一方面，超然性可能被认为作用较小，因为它的主要功能是支持性的（润饰或削弱内容）。最后，NV5:惊奇得到了最多的负面回应（25%）。这可能是因为令人惊讶的标题可能被认为信息量较小，或更加含糊不清。由于人们经常只阅读标题来获取新闻（Ga-bielkov等人，2016），惊讶不会支持标题作为摘要的功能。

总的来说，这项调查的结果强调了新闻价值在标题中的重要性。我们还发现，新闻价值观对英语为母语的人和非母语的人都起着作用（我们的样本中两者的数量大致相等）。这一点很重要，因为现在大多数主要的新闻机构都有更多的全球影响力。

7 结论和未来工作

这里提出的工作是一个更大的项目的第一步

，即利用标题预测新闻文章的受欢迎程度。我们对标题的关注是由于它们在日常在线体验中的作用，其特点是受众注意力有限。

¹³57%的英语为母语的人对“超然”的判断是积极的，而非母语的人只有44%。

以及频繁使用社交媒体网站。

我们提出了一种自动提取*新闻价值*的方法，这种方法已经在新闻研究中被提出，并为数字内容的定性提供了一个新的视角。我们通过开发全自动的、与主题无关的方法来识别标题中的新闻价值，取得了新的突破。一个使用人工注释的评估显示，对于所有的新闻价值，自动提取的输出都与黄金标准相符。众包调查的结果表明，新闻价值影响人们点击标题的决定。这支持在一系列与人类选择有关的应用中更广泛地采用自动分析标题的方法（例如预测模型、推荐系统、智能助手）。

我们目前和未来的工作包括几个阶段。首先，我们已经收集了第二个样本（*纽约时报*）来应用我们的新闻价值提取方法。其次，提取的新闻价值分数正与社交媒体上的头条新闻的流行度相关联，并使用机器学习方法应用于流行度预测模型。人工注释和众包调查的结果也将被用来为预测模型中的特征权重提供参考。此外，另一项调查将针对标题的直接参与度（即读者是否会点击标题），并将其与我们已经准备好的社交媒体流行度指标进行比较。最后，利用众包调查的数据和公开的Twitter数据，我们将研究人口统计学，特别是居住国，是否对突出性和接近性的新闻价值有影响。我们将使用我们从*维基数据*和*巴别网*等知识库中确定的实体数据来丰富这些新闻价值的实现。

鸣谢

这项工作得到了工程和物理科学研究委员会的博士生培训津贴的支持。数据

的收集和存储符合EPSRC的数据管理政策。

该数据集可在

<http://doi.org/10.5518/147>。

我们也要感谢我们的专家注释者的工作和反馈。

参考文献

- Scott L. Althaus, Jill A. Edy, and Patricia F. Phalen. 2001. 在内容分析中使用新闻故事全文的替代物: Which text is best? *American Journal of Political Science*, 45(3):pp.707-723.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. 欧洲语言资源协会 (ELRA)。
- 马可-托莱多-巴斯托斯。2014. 分享、钉子和推特: 从日报到社交媒体的新闻读者群。《新闻研究》, 第1-21页。
- Monika Bednarek 和 Helen Caple。2012. *News Discourse*. Continuum.
- Monika Bednarek 和 Helen Caple。2014. 为什么新闻价值重要? 走向一个新的方法论框架来分析批判性话语分析中的新闻话语及其他。《Discourse & Society》, 25 (2) : 135-158。
- Allan Bell. 1991. *The language of news media*. Blackwell Oxford.
- Jonah Berger 和 Katherine L. Milkman。2012. 是什么让在线内容成为病毒? 《营销研究杂志》, 49 (2) : 192-205。
- 道格拉斯-比伯。1991. *跨越语音和写作的变异*。剑桥大学出版社。
- Jonas Nygaard Blom 和 Kenneth Reinecke Hansen。2015. 点击诱饵: 前瞻性参考作为在线新闻头条的诱饵。《语用学杂志》, 76: 87-100。
- Geert Broëne 和 Seana Coulson. 2010. 处理解放报纸标题中的模糊性: Double grounding. 《话语过程》, 47 (3) : 212-236。
- Helen Caple 和 Monika Bednarek. 2013. 深入探讨话语: 新闻研究中的新闻价值方法及其他。《路透社新闻学研究所》。
- 扬-乔瓦内克。2014. *新闻中时态和时间的语用学: 从典型的头条新闻到在线新闻文本*。Pragmatics & Beyond New Series. 约翰-本-贾明斯出版公司。
- David Conley 和 Stephen Lamb. 2006. 《每日奇迹》: An Introduction to Journalism. 牛津大学出版社。

- 丹尼尔-多尔。2003.论作为相关性优化器的报纸标题。 *Journal of Pragmatics*, 35(5):695-721.
- Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi.2013.内涵词库：表层意义下的一丝情感。In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774-1784, Association for Computational Linguistics.
- Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout.2016.社会点击率：什么和谁在Twitter上得到了阅读？在2016年ACM SIGMETRICS计算机科学测量与建模国际会议论文集中，第179-192页，ACM。
- 约翰-加尔通和玛丽-霍尔姆波-鲁格。1965.外国新闻的结构--四家挪威报纸对刚果、古巴和塞浦路斯危机的报道。 *和平研究杂志*, 2 (1) : 64-90。
- Lorenzo Gatti, Gozde Özdal, Marco Guerini, Oliviero Stock, and Carlo Strapparava.2016.自动创建灵活、吸引人的标题。在 *自然语言处理与新闻学研讨会* 上。IJ-CAI。
- Tony Harcup和Deirdre O'Neill。2001.什么是新闻？Galtung和Ruge的重新审视。 *新闻研究*, 2 (2) : 261-280。
- Karen S. Johnson-Cartee.2005. *News narratives and news framing: Constructing political reality*. Rowman & Littlefield出版公司。
- Hans Mathias Kepplinger和Simone Christine Ehmig。2006.预测新闻决策。对新闻搜索的两部分理论进行了仿真测试。 *通信*, 31 (1) : 25-43。
- Dan Klein 和 Christopher D. Manning.2003.Ac-curate unlexicalized parsing.载于《第41届计算语言学协会年会论文集》，第423-430页，计算语言学协会。
- Onur Kucuktunc, Berkant Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu.2012.对雅虎答案的大规模情感分析。In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 633- 642, ACM.
- Himabindu Lakkaraju, Julian J. McAuley, and Jure Leskovec.2013.名字里有什么？理解社交媒体中标题、内容和社区之间的相互作用。在 *Seventh International AAAI Network Log and Social Media Conference* 上，第311-320页。

萨拉-莱克纳。2012.影响报纸媒体读者阅读行为的展示因素：眼球追踪视角。《视觉传播》，11（2）：163-184。

刘子明。2005.数字环境中的阅读行为：过去十年中阅读行为的变化。《文献杂志》，61(6):700-712。

Rada Mihalcea 和 Andras Csomai. 2007.Wikify!：将文件与百科全书式的知识联系起来。在第十六届ACM信息和知识管理会议论文集上，第233-242页，ACM。

Deirdre O'Neill and Tony Harcup.2009.新闻价值和选择性。《The Handbook of Journalism Studies》，pages 161-174。

Amanda Potts, Monika Bednarek, and Helen Caple.2015.基于计算机的方法如何帮助研究人员调查大型数据集中的新闻价值？《话语与交流》，9（2）：149-172。

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik.1985.《英语的综合语法》。朗文公司。

Marta Recasens, Cristian Danescu-Niculescu-Mizil, 和Dan Jurafsky. 2013.语言学模型的An-对有偏见的语言进行分析和检测。在Proceed-第51届协会年会的会议纪要
《计算语言学》（第1卷：长篇）。
pers), pages 1650-1659, Sofia, Bulgaria, Association for Computational Linguistics.
对斯语、布兰登-奥康纳、丹尼尔-朱拉天斯基和Andrew Ng. 2008. 便宜和快速--但它是评价非专家注解的自然-----。
丰富的语言任务。在2008年的Con-NV2会议上。
自然语言的实证方法会议
处理，第254-263页，美国通信协会。
归纳语言学。

rence Szymanski, Claudia Orellana-Rodriguez, and 特 马克-T.基恩。2016.帮助新闻编辑写作更好的头条：一个推荐者来改善 keyword contents and shareability of news head-行。在自然语言处理遇到Jour-

医学研究会.IJCAI.

陈浩，李丽娟，和彭博。2014.实验结果词语对信息传播的影响：twitter上的主题和作者控制的自然实验。见《计算语言学协会第52届年会论文集》（第一卷：长篇论文），第175-185页，马里兰州巴尔的摩，六月。计算语言学协会。

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer.2003.用循环依赖网络进行富于特征的语篇标记。在《计算语言学协会北美分会2003年人类语言技术会议论文集-第一卷》，第173-180页。

Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos.2004.识别在线搜索查询的相似性、周期性和突发性。在2004年ACM SIGMOD数据管理国际会议上，第2942页。

James G. Webster.2014.《注意力的市场：观众如何在数字时代形成》。麻省理工学院出版社。

AS补充材料：《纽约时报》的专题摘录

表4：《纽约时报》语料库的特征提取统计。表1中解释了符号。报告的措施：中值和最大值，流行率（非零分的比例）。WP=Wikipedia。

特征名称	中位数	最大	患病率
NV1: 突出表现			
实体数量	1	4	100%
可湿性粉剂当前的突发大小	0	57.18	10%
可湿性粉剂的爆裂性	15	166	66%
WP的长期前景	626	65,327	66%
可湿性粉剂的前一天承诺	773	467,458	66%
可湿性粉剂的前一天承诺	0	70	32%
NV2: 情感			
感情	-2	-1	100%
感情	0	1.88	43%
感情	0.25	1	78%
偏见	0.11	1	51%
NV3: 超纲性			
比较/超限	0	1	3%
超限	0	0.33	6%
超限	0	0.33	3%
NV4: 接近性			
靠近	0	1	32%
NV5: 惊喜			
惊喜	4.04	2,724,886	100%

Ori Tenenboim和Akiba A. Cohen.2013.是什么促使用户点击和评论：在线新闻的纵向研究。《新闻学》，16（2）：198-217。

NV6: 独特性

独特性	0	1	34%
-----	---	---	-----
