# 标题很重要：利用标题来预测新闻文章在Twitter和Facebook上的受欢迎程度

**会议论文** - 2017年5月



**4位作者**，包括：

Alicja Piotrkowicz 利兹大学

**24**篇出版物 **177**次引用

查看简介

贾纳-奥特巴赫

塞浦路斯开放大学

**98个**出版物 **2,024次**引用

查看简介

# 标题很重要：利用标题来预测新闻文章在Twitter和Facebook上的受欢迎程度

**Alicja Piotrkowicz**
**Vania Dimitrova**
英国利兹大学计算机学院

**Jahna Otterbacher** 社会信息系统 塞浦路斯开放大学

德国海德堡大学计算机语言学研究所 **Katja Markert** Institut fu¨r Computerlinguistik

**摘要**

像Facebook或Twitter这样的社交媒体已经成为许多读者了解新闻的一个入口。在这种情况下，标题是新闻文章中最突出的部分，而且往往是唯一可见的部分。我们提出了一项新的任务，即*只*使用*标题*来预测新闻文章的受欢迎程度。该预测模型在两个主要的大报新闻机构--《*卫报》*和《*纽约时报》*的标题上进行了评估。我们比几个基线有明显的改进，注意到Facebook和Twitter之间的模型性能差异。

## 简介和相关工作

标题对于吸引读者的注意力和影响他们对新闻的在线阅读体验都是至关重要的。事实上，大约有六分之一的人在阅读时只看标题，而不点击文章全文的链接[1]。眼球追踪研究从经验上证实了这种行为；许多人是 "入门级读者"，他们关注标题以确定文章的概况，但他们的阅读活动很少（Holsanova, Rahm, and Holmqvist 2006）。此外，在许多网络空间中，标题是新闻文章中唯一可见的部分；例如新闻提要和社交媒体。

然而，尽管如此，头条新闻并没有被认为是新闻文章受欢迎程度预测的唯一数据来源。大多数模型利用了发布后的数据，比如早期采用者的数量（Castillo等人，2014）。这些方法对*人气的发展进行*建模，例如，他们可能使用*文章发表后*第一小时内的推文数量来预测后期或最终的人气。

另一方面，解决Ara-pakis、Cambazoglu和Lalmas（2014）所说的 "冷启动问题 "的方法，即*在发表前*预测新闻文章的受欢迎程度，仍然处于起步阶段。特别是，这些方法对新闻文章文本的哪些方面使其在网上受欢迎提供了有限的洞察力。Bandari、Asur和Huberman（2012）使用了少量与主题类别、命名实体的突出性和情感有关的文本特征。Arapakis、Cambazoglu和Lalmas（2014）在Bandari、Asur和Huberman的工作基础上进行了重新制作和改进。



[1]http://bit.ly/21LwfS5

Huberman（2012）。他们还增加了少量的语言学和突出性特征，但他们主要关注的是评估方法。Bandari, Asur, and Huberman (2012)和Arapakis, Cambazoglu, and Lalmas (2014)都将新闻来源作为一个特征，这被证明是受欢迎程度的压倒性决定因素。然而，如果新闻编辑部的工作人员想调整文章内容以获得更多的受众，这是无帮助的，因为新闻来源是他们无法控制的。此外，以前的这些模型都是把标题和文章正文放在一起考虑。由于标题在网络新闻领域发挥着至关重要的作用，因此值得研究的是，我们能在多大程度上仅从标题上预测一篇文章的受欢迎程度。我们的目标是研究从标题中提取的各种文本特征，并确定它们是否对新闻文章的社会媒体流行度有影响。我们通过以下方式加强先前的工作：(i) 仅使用标题；(ii) 引入新的特征；(iii) 使用源内部评估。

## 数据收集

我们创建了两个新闻头条语料库，并获得了每个头条的社会媒体流行度。

**新闻机构。**我们使用两个主要的大报--*《卫报》*和*《纽约时报》*。我们下载了2014年4月（卫报训练）、2014年7月（卫报测试）、2014年10月（纽约时报训练）和2014年12月（纽约时报测试）期间发布的所有标题[2]。表1包括了一些标题的例子和它们的受欢迎程度的分数。

**社会媒体数据。**我们通过一篇新闻文章在Twit-ter和Facebook上被引用的次数来衡量它的社会媒体流行度。文章的URL被用作Twitter搜索API的搜索查询[3]，以获得文章发表后一天、三天和七天的推特和转发次数。使用Facebook FQL API对Facebook的喜欢和分享重复这一过程。[4]

**受欢迎程度的衡量。**推文和转发，以及分享和喜欢，被合并为两个指标：推特和脸书的人气。我们发现，在我们的数据集中，Twitter和Facebook在三天和七天后的受欢迎程度并无显著差异，因此在整个文件中，我们报告了

[2] *卫报》的*数据：监护人内容API，*纽约时报*：NYT文章搜索API。

[3] https://dev.twitter.com/docs/api/1.1/get/search/tweets

[4] https://developers.facebook.com/docs/technical-guides/fql/

表1：最受欢迎和最不受欢迎的标题的例子。

| | *卫报* | *纽约时报* |
|---|---|---|
| 最受欢迎的 | "资本主义根本不可行，原因在此"（T=2299，F=23840）。 | "纽约市的医生得了埃博拉病" (T=12780, F=46603) |
| 最不受欢迎 | "更正和澄清" (T=0, F=0) | "Pastis and Ouzo: The Soccer of Liquors" (T=0,) " |
| | 对让领主辞职的担心是错误的" (T=5, F=0) | "阿拉斯加的政治前景" (T=0, F=1 |

三天后的受欢迎程度，产生了两个社会媒体的流行度：T = 三天后的Twitter人气，F = 三天后的Facebook人气。

**数据概述。**受欢迎程度显示出强烈的Zipfian分布。Twitter和Facebook的衡量标准相互之间有很好的相关性（Guardian: $\rho$=0.74, NYT: $\rho$=0.6）。然而，Twitter显示出比Facebook更平坦的分布。在两个数据集中，Facebook的引用次数都比Twitter高得多，这可能是由于用户数量的原因（2016年，Facebook有17亿活跃用户，而Twitter只有3亿[5]）。新闻来源也发挥了重要作用，因为*《纽约时报》*的文章更经常在社交媒体上分享（这遵循了Bandari、Asur和Huberman（2012）的发现，即新闻来源是预测新闻文章在社交媒体上受欢迎程度的最强因素）。

## 标题特点

我们使用两种类型的特征：受新闻学启发的*新闻价值*和*语言学风格*。特征实现的细节在Piotrkowicz, Dimitrova, and Markert（2017）中概述。

### 新闻价值

新闻价值是一个源于新闻学研究的概念，指的是新闻故事中使其具有新闻价值的方面。虽然有许多新闻价值分类法，但有相当多的重叠（参见Caple和Bednarek（2013）），我们实现了六个经常包括的新闻价值。

**突出性。**对突出实体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所获得的在线关注量。我们通过使用维基化获取实体来扩展以前的工作，这确保了实体类型的广泛性。我们实现了六个 "突出性 " 特征：(i)维基化实体的数量；(ii)新闻显著性（实体在相关新闻媒体头条中被提及的数量）；(iii)长期显著性（一个实体在一年中每天维基百科页面浏览量的中位数）；(iv)前一天的突出性（某实体在特定标题发布前一天的维基百科页面浏览量）；(v)当前的爆发性大小（如果某实体是 "爆发性的"，那么该实体的页面浏览量比平均水平高出多少）；以及(vi)爆发性（如果某实体是 " 爆发性的"，那么该实体在一年内有多少次的爆发性）。

(*maxPos+maxNeg*)；(iii)有偏见的词的比例；和(iv) 正面/负面含义的词的比例。

**规模。**这是指一个新闻事件的规模或影响。有三个特征：(i) 比较级和最高级词汇的比例（基于POS 标签）；(ii) 强化词的比例；以及(iii) 下降词的比例。

**接近性。**我们专注于与新闻来源的地理接近性，这假定来自与新闻机构相同国家的读者构成其读者群的很大一部分。我们将Proximity实现为标题文本中对新闻出口国家（英国/ 美国相关关键词）的明确提及。

**惊喜。**令人惊讶的标题会吸引人们的注意。我们参照维基百科语料库[6]，通过计算标题中的句法块的共同性来衡量惊喜。**独特性。**标题应该是新颖的。为了研究这一点，对于一个给定的标题，我们通过最近的过去的标题，看看是否有任何高度相似的标题。对于一对标题和过去的标题向量（使用*tf-idf*加权的Gigaword语料库创建），我们计算余弦相似度。最高的余弦相似度被指定为特征值。

**感情。**情绪指的是负面和正面。它的词汇量。我们使用Senti- WordNet（Baccianella, Esuli, and Sebastiani 2010）的分数来计算四个特征：(i) 感情（*maxPos - maxNeg - 2*）；(ii) 极性

[5]http://bit.ly/2ddRJHi

## 风格

对于语言风格，我们计算了受期刊研究和NLP工作启发的关于措辞效果的特征。

**简明扼要。**标题需要简短。我们将简洁性作为标记的数量和字符的数量来实现。

**简单性。**易于理解的标题使用简单的同义词和词汇。我们使用两个句法复杂度指标：(i) 解析树的高度，和(ii) 非终端树节点的数量。为了测量词汇的复杂性，我们采用了四个特征：(i) 熵（使用CMU-Cambridge Toolkit 在Gigaword 语料库的*纽约时报*部分建立的八卦语言模型计算）；(ii) 困难词的比例（任何不出现在语言模型中5000 个最常见的词中的词）；(iii) 绵密的词频（使用未对称的词频列表[7]）；(iv) 信息含量（在英国国家语料库中计算名词和动词）。

**毫不含糊。**新闻文本不应该有歧义。我们使用两个特征来衡量标题的模糊性：WordNet中每个词的中位数，以及模态（模态事件或事件间的模态关系；使用TARSQI[8]）。

**标点符号。***卫报》*的标题风格指南[9]

[6]http://www.nlp.cs.nyu.edu/wikipedia-data

[7]http://www.wordfrequency.info/；卫报的英国国家语料库，纽约时报的当代美国英语语料库。

[8]http://www.timeml.org/site/tarsqi/toolkit/index.html

[9]http://www.theguardian.com/guardian-observer-style-guide-h

不鼓励使用引号、问号和感叹号。我们实现了表明它们存在的二进制特征。

**名词。**《*卫报*》的风格指南告诫我们不要使用过多的连续名词（所谓的 "headlinese"）。我们实现了四个特征：(i) 三个连续的名词（双名词特征）；(ii) 名词短语的数量；(iii) 普通名词的比例；以及(iv) 专有名词的比例。

**动词。在**《*卫报*》的风格指南中，鼓励在标题中使用动词。我们实现了两个特征：(i) 动词短语的数量，和(ii) 动词的比例。

**副词。**副词，尤其是方式副词，经常被用于标题中。我们使用副词的比例。

## 预测模型

利用这些特征，我们的目标是通过文章的标题来预测Twitter和Facebook上的新闻文章的受欢迎程度。我们为每个新闻来源建立单独的预测模型，从而避免了新闻来源的流行效应。

### 方法

我们使用了回归法（支持向量回归法与RBF卡特尔）。Arapakis、Cambazoglu和Lalmas（2014）认为，使用分类法进行人气预测并不合适，因为类的分割可能会对低人气的文章带来偏见。流行度量--T、F--是经过对数转换的，以提高模型的适应性。

在测试集[10]，对结果进行了评估。使用了两个评价指标：Kendall's tau等级相关系数（$\tau$）和平均绝对误差（MAE）。对$\tau$的显著性检验采用$z$*检验*，对MAE的显著性检验采用$t$*检验*。

### 基线

我们使用了三个基线：一个单字基线和两个最新的基线。我们的模型的特征表示为$M$. **单词（$M_U$）。**我们使用了1000个最频繁的单字。**最先进的重新实施：**Bandari, Asur, and Huberman (2012) ($M_B$)和Arapakis, Cambazoglu, and Lalmas (2014) ($M_A$)最初使用了完整的文章文本，但我们在与我们相同的数据集上运行这些基线（即只有标题行）。我们的目标是尽可能地重新实现，但在某些情况下我们不得不进行调整。我们使用了斯坦福命名实体识别器和SentiWordNet，分别用于突出性和情感特征。由于无法获得Twitter的存档数据，我们使用维基百科来计算突出性特征。最后，与原来的实现不同的是，没有新闻来源特征（因为我们的目标是是一个源头内部的评价）。

这两种最先进的方法以及类似的任务（Lakkaraju, McAuley, and Leskovec 2013）都利用了文章发表时的元数据（类别、时间）。重新实施的基线和我们的完整模型（$M$）也包括元数据。按照Arapakis、Cambazoglu和Lalmas（2014）的实现，类别和发表日期和时间都被实现为

[10]使用交叉验证法进行评估是不合适的，因为数据是有时间顺序的，而我们的特征之一，标题的唯一性，是利用了时间顺序的。

在我们的模型中的二元特征。Bandari, Asur, and Huberman (2012)计算了一个类别得分 ( *#引用次数 per 类别*)。

## 结果和讨论

我们报告了针对不同基线的回归结果。

表2：基线与我们模型的回归结果
(*M*)使用所有特征（新闻价值、风格、元数据）。粗体字的结果表示改进程度为*P<0.*05。

| | *卫报* | | | | *纽约时报* | | | |
|---|---|---|---|---|---|---|---|---|
| | τ | | MAE | | τ | | MAE | |
| | T | F | T | F | T | F | T | F |
| *MU* | 0.32 | 0.25 | 0.82 | 1.59 | 0.19 | 0.22 | **0.66** | 1.68 |
| *MB* | 0.36 | 0.29 | 0.71 | 1.53 | 0.15 | 0.18 | 0.67 | 1.72 |
| *马云* | 0.41 | 0.35 | 0.7 | 1.45 | 0.21 | 0.3 | 0.86 | 1.57 |
| ***M*** | **0.43** | **0.37** | **0.68** | **1.42** | **0.23** | **0.32** | 0.88 | **1.54** |

表3：基线与我们模型的回归结果
(*M*) 仅使用标题特征（新闻价值和风格）。粗体字的结果表示改进程度为*P<0.*05。

| | | | | | *卫报》纽约时报* | | | |
|---|---|---|---|---|---|---|---|---|
| | τ | | MAE | | τ | | MAE | |
| | T | F | T | F | T | F | T | F |
| $M_B$ | 0.11 | 0.07 | 0.94 | 1.74 | 0.05 | 0.02 | 0.7 | 1.85 |
| $M_A$ | 0.22 | 0.19 | 0.88 | 1.66 | 0.19 | 0.16 | **0.67** | 1.75 |
| ***M*** | **0.29** | **0.26** | **0.83** | **1.59** | **0.21** | **0.23** | 0.69 | **1.66** |

**使用全部特征集与基线的性能（表2）。**我们的模型（*M*）在几乎所有的衡量标准上都明显优于基线。例外的情况是纽约时报数据集中Twitter的MAE结果，其中uni- grams基线优于该模型。然而，对于同一数据集，我们的模型取得了明显更高的关系。在《卫报》数据集中，Twitter取得了最好的结果（τ=0.43，MAE=0.68）。这是一个很有希望的结果，考虑到这是第一次尝试使用头条来预测新闻文章的受欢迎程度。

**与仅使用内容特征的基线相比的性能（表3）。**当局限于可直接从标题文本（新闻价值和风格）中提取的特征时，我们的模型在大多数衡量标准上都比使用元数据的模型有相当大的改进（相关度的改进约为40%，而使用元数据时为5-10%）。在《卫报》的数据集中，Twitter的相关度最高（τ=0.29），而在纽约时报的数据集中，Twitter的MAE最低（MAE=0.69）。

**特征组的性能（表4）。**在*P<0.*01的情况下，使用所有特征明显优于任何单独的特征组。同样，例外的是纽约时报数据集中Twit- ter的MAE结果。虽然新闻值达到了

*每个类别的#文章*

表4：比较特征组的回归结果（$M_N$ = 新闻价值，$M_S$ = 风格，$M_M$ = 元数据）。粗体字的结果表示改进程度为$P<0.01$。

| | 卫报 | | | | 纽约时报 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\tau$ | | MAE | | $\tau$ | | MAE | |
| | T | F | T | F | T | F | T | F |
| 明尼苏达州 | 0.2 | 0.17 | 0.89 | 1.67 | 0.14 | 0.14 | **0.68** | 1.74 |
| $M_S$ | 0.25 | 0.22 | 0.86 | 1.62 | 0.18 | 0.19 | 0.7 | 1.7 |
| $M_M$ | 0.39 | 0.33 | 0.72 | 1.51 | 0.17 | 0.23 | 0.92 | 1.65 |
| $M$ | **0.43** | **0.37** | **0.68** | **1.42** | **0.23** | **0.32** | 0.88 | **1.54** |

在所有组别中性能最低，与Twit-ter和Facebook流行度的相关度仍在0.14和0.2之间。特别值得注意的是，在很大程度上与主题无关的风格特征，其本身就取得了良好的表现（高达0.25的相关性和0.7的MAE）。这表明，标题风格对社交媒体读者很重要，与文章内容无关。这似乎遵循了之前关于在线内容流行度预测的研究，其中风格的各个方面也被发现对流行度有影响（Tan, Lee, and Pang 2014）。元数据（特别是类别）取得了良好的效果，特别是对于《卫报》的数据集，表明文章的主题和体裁对读者起着重要作用。尽管元数据增加了预测性能，但应该注意的是，新闻文章的这一方面通常不是由作者控制的（也就是说，人们不能轻易改变体裁或主题）。另一方面，大多数新闻价值和风格特征可以自由编辑，以达到更高的人气。

**Twitter和Facebook之间的差异。**也许是由于分布更加偏斜，Facebook的误差更大。对于相关性，Twitter在卫报的数据上表现得更高，而对于纽约时报则相反。Facebook和Twitter上不同的新闻读者人口统计学[11]，这可能是原因之一，这需要进一步的工作，考虑到用户人口统计学。

**新闻来源之间的差异。**我们工作的一个关键方面是来源内部的评估，这在以前的工作中是没有的。事实上，《卫报》的数据表现比《纽约时报》好。这指出了在其他新闻机构和类型（例如小报）方面的进一步工作。

**计算成本与性能。**当与表现最好的基线（$M_A$）相比较时，我们的完整模型取得了明显的改进（参见表2）。当只考虑直接从标题文本中提取的特征时，这种差异就更加明显了（参见表3）。虽然使用所有可用的特征的整体性能只比最先进的模型略有提高，但使用可以被标题作者更容易编辑的特征（可能会增加其受欢迎程度）的模型显示出相当大的改进。

[11]http://pewrsr.ch/27TOfhz

## 总结

新闻标题在社交媒体上发挥着至关重要的作用。在一在使用标题提取的特征来预测新闻文章在社会媒体上的受欢迎程度的新任务中，我们比几个基线有明显的改进。从标题文本（通常可由标题作者编辑）中提取的特征在单独考虑时对预测性能有影响。这表明，传统的编辑对新闻价值的判断和NLP研究对风格的洞察都适用于预测社交媒体上的标题流行度。我们的特征提取方法是通用的，可以在不同的新闻发布和体裁中重复使用。预测模型的结果取决于新闻来源；进一步的工作可以包括在不同的新闻渠道和在线内容中进行性能比较。我们目前正在完善预测模型，考虑到用户的人口统计学，并整合世界知识。首先，我们正在考虑用户的位置（居住国），以改善邻近性特征。其次，为了改善突出性（我们的最佳相关特征），我们正在从维基数据中整合世界知识，以关联实体对用户的位置有意义。

## 鸣谢

这项工作得到了英国EPSRC的博士培训津贴的支持。数据收集和存储符合EPSRC的数据管理政策。该数据集可在https://doi.org/10.5518/174。

## 参考文献

Arapakis, I.; Cambazoglu, B. B.; and Lalmas, M. 2014。关于在冷启动时预测新闻流行度的可行性。In *SocInfo*, 290-299.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010.Sentiwordnet 3.0：用于情感分析和意见挖掘的强化词汇资源。在*LREC*。

Bandari, R.; Asur, S.; and Huberman, B. A. 2012。社交媒体中的新闻脉搏：预测人气。在*ICWSM中*。Caple, H., and Bednarek, M. 2013.深入研究社会化媒体中的课程：新闻研究中的新闻价值方法及其他。*路透社新闻学研究所*。

Castillo, C.； El-Haddad, M.； Pfeffer, J.； 和Stempeck, M. 2014。使用社交媒体的反应来描述在线新闻故事的生命周期。在*CSCW*。

Holsanova, J.; Rahm, H.; and Holmqvist, K. 2006。报纸散页上的切入点和阅读路径：比较符号学分析和眼球追踪测量。*视觉交流*5（1）：65-93。

Lakkaraju, H.; McAuley, J. J.; and Leskovec, J. 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*ICWSM*。

Piotrkowicz, A.; Dimitrova, V. G.; and Markert, K. 2017。从标题文本中自动提取新闻价值。在*EACL 2017学生研究研讨会上*。

Tan, C.; Lee, L.; and Pang, B. 2014。措辞对信息传播的

影响：在Twitter上进行的主题和作者控制的自然实验。在*ACL*。