信息处理与管理57(2020)102034

ScienceDirect提供内容列表

信息处理与管理

期刊主页： www.elsevier.com/locate/infoproman

对两个在线社交网络 Twitter 和 Reddit 中的文档聚类和主题建模的评估

Stephan A. Curiskis ⁎ Barry Drake、Thomas R. Osborn、Paul J. Kennedy

悉尼科技大学工程与信息技术学院人工智能中心,15 Broadway,Ultimo,新南威尔士州
2007年澳大利亚

文章信息

关键词：
文档聚类
主题建模
主题发现
嵌入模型
在线社交网络

抽象的

在线社交网络 (OSN) 中的文档聚类和主题建模方法提供了一种对大量用户生成的内容进行分类、注释和理解的方法。
多年来已经开发了许多技术,从文本挖掘和聚类方法到潜在主题模型和神经嵌入方法。然而,这些方法中的许多方法在应用于 OSN 数据时效果不佳,因为此类文本非常短且噪音大,而且研究结果通常不具有可比性。在本研究中,我们在 Twitter 和 Reddit 的三个数据集上评估了几种文档聚类和主题建模技术。我们对从词频逆向文档频率(tf-idf)矩阵和词嵌入模型与四种聚类方法相结合得出的四种不同特征表示进行了基准测试,并且我们包括一个潜在狄利克雷分配主题模型进行比较。

文献中使用了几种不同的评估方法,因此我们针对此任务最合适的外部评估方法进行了讨论并提出了建议。我们还展示了这些方法在不同文档长度的数据集上的性能。我们的结果表明,使用适当的外部评估方法,应用于神经嵌入特征表示的聚类技术在所有数据集上均能实现最佳性能。我们还展示了一种使用基于热门词的方法解释聚类的方法,该方法结合了tf-idf权重和嵌入距离测量。

一、简介

2018 年 1 月,全球估计有 40.21 亿人使用互联网。其中,31.96 亿人以某种形式使用社交媒体,产生了惊人的内容量。[1]在线平台和社交网络已成为近一半世界人口的主要信息来源。这些平台越来越多地被用于传播有关新闻、品牌、政治讨论、全球事件等的信息（Bakshy、Rosenn、Marlow 和 Adamic,2012 年）。然而,生成的大部分数据都是非结构化的,没有注释。这意味着很难理解信息主题是如何通过在线社交网络 (OSN) 传播的,以及用户如何参与不同的主题（Guille、Hacid、Favre 和 Zighed,2013 年）。自动注释 OSN 中的主题可以通过丰富这些平台提供的数据来促进信息传播和用户偏好的分析,从而易于分析。随着回音室和过滤气泡等现象的兴起,个人会收到带有偏见和狭隘内容,

⁎ 通讯作者。
电子邮件地址： stephan.a.curriskis@student.uts.edu.au （SA 库里斯基斯）。

[1] https://wearesocial.com/uk/blog/2018/01/global-digital-report-2018， 2018 年 9 月访问。

https://doi.org/10.1016/j.ipm.2019.04.002 2018
年 9 月 22 日收到； 2019 年 4 月 11 日收到修订版； 2019 年 4 月 11 日接受2019 年 4 月 17 日在线发布 0306-4573/

自动注释 OSN 数据变得很重要。

文档聚类是一组机器学习技术,旨在自动将文档组织到集群中,以便与其他集群中的文档相比,集群内的文档相似。人们已经提出了许多聚类文档的方法（Bisht & Paul,2013; Naik,Prajapati,& Dabhi,2015）。这些技术通常涉及使用特征矩阵（例如术语频率逆文档频率矩阵（tf-idf矩阵））来表示语料库,并对该矩阵应用聚类方法。最近,源自神经词嵌入的表示已在社交媒体数据上得到应用,因为它们可以产生具有语义属性的密集表示,并且比传统方法需要更少的手动预处理（Li、Shah、Liu 和 Nourbakhsh,2017）。在这种情况下应用的常见聚类方法会构建层次结构或分区（Irfan 等人,2015）。分层方法的示例是凝聚聚类和分裂聚类。示例划分方法是 k-means 和 k-medoids 聚类。

主题建模涉及发现文档中词语使用模式的方法,这是一个活跃的研究领域,最近有几种技术应用于 OSN 数据（Chinnov、Kerschke、Meske、Stieglitz 和 Trautmann,2015 年）。主题通常定义为词语的分布,而文档则被建模为主题的混合。与文档聚类一样,主题建模可用于通过为每个文档提供一系列主题的概率分布来对文档进行聚类。这可以看作是一种软分区聚类,其中数据点对每个聚类具有概率所有权。主题表示还提供了每个主题的词语分布,有助于解释。在 OSN 文本数据上应用的常用主题模型包括潜在狄利克雷分配（Blei、Ng 和 Jordan,2003 年）、作者主题模型（Hong 和 Davison,2010 年)以及最近的动态主题模型,该模型可随时间发现主题（Alghamdi 和 Alfalqi,2015 年）。

文档聚类和主题建模是越来越重要的研究领域,因为这些方法可以应用于大量现成的 OSN 文本数据,从而产生同质的文档组。然后,这些文档组可能与相关主题和趋势保持一致。聚类特别适合 OSN 数据,因为 Twitter 和 Facebook 等平台使用主题标签作为主题注释的一种形式（Steinskog、Therkelsen 和 Gambäck,2017）,可用于评估文档聚类和主题建模方法。大规模聚类有助于理解每天在线创建的大量内容,随后可用于进一步的机器学习任务。从 OSN 数据（例如用户人口统计、地理和网络数据)得出的其他特征也被聚类以查找语义相似的在线帖子或评论组（Alnajran、Crockett、McLean 和 Latham,2017）。然而,在应用主题建模和文档聚类方法时,OSN 数据带来了许多挑战。例如,这样的文本通常很短,并且包含拼写错误和语法错误等噪音（Chinnov 等,2015）。

OSN 数据集上的主题建模和文档聚类研究面临两个关键挑战。首先,结果通常不可重现,因为研究中使用的数据通常无法发布。例如,Twitter 的服务条款不允许发布推文。相反,研究人员可以发布通过 API 使用和检索的推文标识符列表。

不幸的是,随着时间的推移,相关推文会从平台中删除,这会降低底层数据的质量。使用的数据集通常也很小或偏向于特定的上下文。这些问题是由于从 OSN 平台提取大型数据集通常需要复杂的数据收集和准备以及平台本身的限制造成的（Stieglitz、Mirbabaie、Ross 和 Neuberger,2018）。

其次,不同的研究通常使用不同的方法来评估聚类文档的性能。 Twitter 数据的评估方法多种多样,从将集群与标记数据进行比较的外在测量,到集群性能和可解释性的手动评估（Alnajran 等人,2017）。因此很难比较经验结果。随着该领域研究的快速发展,对于哪种方法或方法系列在特定情况下（例如较短的 Twitter 数据或相对较长的 Reddit 评论)表现最佳的指导很少。

在本文中,我们分析了三种数据集上几种用于 OSN 内容文档聚类和主题建模的方法的性能:两个 Twitter 数据集和一个公开的 Reddit 数据集。我们评估了四种特征表示方法,这些方法源自tf-idf和嵌入矩阵,并结合了四种聚类技术,并包括一个潜在狄利克雷分配 (LDA) 主题模型以供比较。我们还讨论了文献中常用的文档聚类评估指标的属性和适用性。我们使用三种指标来评估性能,即标准化互信息 (NMI)、调整后的互信息 (AMI) 和调整后的兰德指数 (ARI)。此外,我们提供了我们的数据集,以便可以重现我们的结果。为了遵守 Twitter 的使用条款,我们提供了与主题标签一起使用的推文标识符。我们还提供了使用的完整 Reddit 数据集（Curiskis、Drake、Osborn 和 Kennedy 已提交）。

此外,通过调整关键超参数,我们展示了如何使用嵌入模型生成文档聚类的特征集,从而提供良好的性能并捕获数据中的潜在结构。我们还展示了单词嵌入距离如何通过对顶级单词进行排序来协助解释聚类,从而形成单词的主题向量。这一贡献非常重要,因为 OSN 的数据集通常很短,并且包含诸如拼写错误、缩写、首字母缩略词、特殊字符、表情符号、URL 和主题标签等噪音。这些问题可能会导致许多常用技术的性能不佳。

此外,关于有效处理 OSN 数据的方法,文献中缺乏明确的共识。本文的结果为在不同类型的 OSN 数据上提供良好性能的方法提供了指导。这些结果表明,传统的主题建模和文档聚类方法在简短且嘈杂的社交媒体帖子上效果不佳。相反,应用于更新的神经网络嵌入表示的聚类方法可以提供改进的性能。

本文的结构如下。第 2 节回顾了该研究领域的现有文献。第3 节介绍了我们的方法的细节,包括数据提取、准备过程、特征表示、聚类方法和评估指标的描述。第4 节介绍了我们的结果并进行了讨论。第5 节提供了

讨论之后,我们将在第 6部分得出结论。

2。文献综述

我们将有关 OSN 的文档聚类和主题建模的文献分为三个领域。首先,许多研究都集中在识别和解释该领域的模因,结合文本、网络和用户数据。其次,通过主题模型和聚类方法识别主题作为理解和分类在线内容的手段受到了广泛关注。第三,神经词嵌入模型的最新进展已被用于提供来自 OSN 的文档的密集特征表示。

2.1.模因识别

“模因”一词通常用于表示通过模仿从一个人传播到另一个人的文化元素或行为系统。在 OSN 的背景下,在本文中,我们将“模因”定义为以电子文本表示的语义单元,其中语义在多个个体之间传递,即使文本可能不同。“模因”的这种特定定义有时被称为“ememe” （Shabunina & Pasi,2018）。OSN应用程序中的主题可以定义为一组连贯的语义相关术语,它们表达一个论点（Guille 等人,2013）。与主题的这种定义相比,模因不一定需要从一组或一组单词中得出,而是旨在检测重要的语义内容。然而,在实践中,这两个概念往往存在重叠。模因的概念对于 OSN 应用程序很有用,因为它可以被视为文本内容的潜在表示,但也可以通过分析 OSN 用户和网络数据来发现。

Ferrara 等人 (2013)的一项研究旨在识别大型社交媒体数据中的 meme。在该研究中,为 Twitter 数据定义了几种相似度度量,这些度量利用了内容、元数据和网络特征。作者定义了“pro-tomeme”的概念,用于指代主题标签、用户提及、URL 和短语。通过基于推文、用户和内容特征在空间上创建 protomeme 投影来聚合数据。对于每个 protomeme 对,计算共同的用户、推文、内容和传播相似度度量。然后以几种不同的方式聚合这些相似度矩阵,例如元素平均值和最大值。最后,使用层次聚类对聚合相似度矩阵进行聚类。得到的聚类用于表示数据中的 meme。使用的数据集是与 2012 年 4 月美国总统初选相关的 5523 条推文的集合。手动识别了 26 个主题并将其作为标签分配给每条推文。由于每条推文中的模因和主题可能会重叠,因此使用标准化互信息变体（称为 LFK-NMI)来评估性能。考虑到此方法的最佳参数,原始模因聚类方法的平均 5 倍交叉验证 LFK-NMI 得分约为 0.13。JafariAsbagh 、Ferrara、Varol、Menczer 和 Flammini（2014 年)后来扩展了该算法以处理流数据。

最近， Shabunina 和 Pasi（2018)开发了一种识别和表征模因的方法,模因被视为一组随着时间的推移通过网络传播的频繁出现的相关单词。社交媒体流中术语之间的关系是使用单词图来建模的。为了识别迷因,对图应用了 k 核简并过程来生成子图,这些子图构成了迷因基础。模因被定义为模因基础中术语的模糊子集。该方法已应用于来自搜索查询 #economy、#politics 和 #finance 的超过 800,000 条推文。尽管模因对于描述和解释社交媒体流中的主题很有用,但它并不归因于个人社交媒体文档或用户。该方法的评估仅限于主观解释和内在措施。

2.2.文档聚类和主题建模

与 meme 识别方法不同,许多研究都专注于检测 OSN 中的主题。主题模型通常指对包含相似单词的文档以及出现在相似文档集中的单词进行分组的方法。

文档聚类是指根据某些特征矩阵对文档进行分组的方法,这样集群中的文档与其他集群中的文档更相似。由于文档较短且 OSN 数据（如 Twitter 数据)固有噪声程度较高,因此通常采用基于聚类的方法代替更传统的主题模型（Chinnov 等人,2015 年）。尽管如此,应用于 OSN 数据的主题模型仍然是一个活跃的研究领域（Alghamdi 和 Alfalqi,2015 年）。事实上，“主题发现”一词可能指主题建模或文档聚类。

文档聚类方法通常使用文档中单词出现的向量空间表示。通常,词袋方法将每个文档建模为词空间中的一个点。每个单词都是该空间的一个特征或维度,元素值以多种方式之一分配。这些可以是 one-hot-encoding,其中如果文档中存在该单词,则该值设置为 1,否则设置为 0、术语频率或术语频率逆文档频率计算。鉴于总维度大小是唯一单词的数量,通常存在仅使用那些具有高值的单词的阈值（Patki&Khot,2017）。然后可以将一系列聚类算法应用于特征矩阵,例如 k 均值、层次聚类、自组织映射等（Naik 等人,2015）。

例如， Godfrey、Johns、Meyer、Race 和 Sadek (2014)开发了一种算法来识别特定 Twitter 数据集中的主题,该数据集是使用查询词 “世界杯”提取的约 30,000 条推文的集合。将非负矩阵分解 (NMF) 和 k 均值聚类应用于推文的tf-idf表示以创建主题聚类。由于 Twitter 数据的噪声性， Godfrey 等人 (2014)开发了一个初步过滤步骤,使用多次运行 DBSCAN 聚类算法结合共识聚类。其原理是,不接近任何特定聚类的推文可能会被视为噪声并从分析中删除。使用此方法的结果表明,k 均值聚类和 NMF 都产生了相似的结果。

然而,当使用推文网络图和词云的主观评价来分析聚类时,NMF 似乎产生了更多可解释的聚类。

Fang、Zhang、Ye 和 Li (2014)使用有关推文的附加信息来检测 Twitter 中的主题。认识到推文的文本内容可能非常有限,因此开发了一个基于更细粒度的“多关系”的“多视图”主题检测框架。这些多关系被定义为来自 Twitter 社交网络的有用关系,包括主题标签、用户提及、转发、有意义的单词和相似的发布时间。为了测量这些多关系,开发了一个文档相似性度量。然后将多关系相似性分数组合成多视图并使用三种不同的方法进行聚类。这些聚类被用来表示主题,并应用基于后缀树和tf-idf权重的关键字提取方法来为每个聚类得出代表性关键字。使用从 Twitter API 中提取的 12,000 条推文数据集(其中包含 60 个“热门”主题)对该方法进行了评估。使用了三种评估指标,即F 度量、NMI 和熵。

结果表明,纳入更多多视图可以提高性能,F 值结果高于 0.928 ,NMI 高于 0.935。不过,作者并没有从文本中删除任何热门话题关键词。这些关键词通常是短语或主题标签,可以通过tf-idf方法轻松发现。

另一项研究比较了不同聚类方法检测 Twitter 数据中以尼泊尔近期地震为中心的主题的效果(Klinczak & Kaestner,2016)。在这项研究中,推文由它们的tf-idf向量表示。比较了应用于此表示的四种聚类方法,即 k-means、k-medoids、DBSCAN 和 NMF。通过使用簇的内聚性和分离性度量(即内在评价度量)来评估每种聚类方法,很明显,NMF 产生了更简单且更易于解释的优质簇。最近, Suri 和 Roy (2017)应用 LDA 和 NMF 来检测 Twitter 数据集以及 RSS 新闻源上的主题。发现两种方法具有相似的性能。 LDA 被认为更具可解释性,但 NMF 的计算速度更快。然而,性能是通过手动检查主题的关键术语来评估的。

许多研究已将主题建模技术应用于 OSN 数据。例如, Paul 和 Dredze (2014)开发了一个主题建模框架,用于使用 Twitter 数据发现自我报告的健康主题。5128 条推文被标注为积极状态,如果它们与用户的健康有关,则标注为消极状态。训练逻辑回归模型来预测标注数据中的积极标签,并将其应用于经过大量健康相关关键词过滤的 Twitter 流。这提供了一组 1.44 亿条健康推文,用于运行疾病主题方面模型。虽然这项研究有助于过滤和解释大量相关推文,但对发现的主题的验证侧重于与外部健康趋势数据的相关性测量。

除了应用于静态数据集的主题模型之外,结合了 OSN 数据的时间性质的动态主题模型也越来越受到关注(Alghamdi & Alfalqi,2015)。 Ha、Beijnon、Kim、Lee 和 Kim (2017)将动态主题模型应用于 Reddit 数据,以了解用户对智能手表的看法。虽然这些结果对于衡量该领域的公众舆论很有趣,但没有使用真实标签,同样也没有应用外在评估措施。最近, Klein、Clutton 和 Polito (2018)应用主题建模来揭示 Reddit 阴谋页面(Reddit 子页面)中的不同兴趣。 NMF 用于为每个对页面做出贡献的用户创建主题加载。然后使用 k 均值对这些主题负载进行聚类以揭示用户子组。同样,这项研究对于了解 OSN 讨论线程中的用户群体很有用,但没有进行外部评估来验证主题建模或聚类的质量。

2.3.神经网络嵌入模型

关于 OSN 文本数据聚类的许多文献在某种程度上使用了推文的tf-idf矩阵表示。这些矩阵将术语视为独热编码向量,其中每个术语由一个二进制向量表示,该向量恰好有一个非零元素。这意味着单词之间的关系(例如同义词)不会被纳入,并且生成的文档矩阵表示是稀疏且高维的。密集的分布式单词表示或词向量嵌入的概念提供了一种替代方法(Bengio、Ducharme、Vincent 和 Janvin,2003 年)。在这些方法中,每个单词都由一个固定维度的实值向量表示。词向量通常使用神经网络语言模型进行训练,例如 word2vec (Mikolov、Chen、Corrado 和 Dean,2013 年)。但是,当使用词向量嵌入模型创建文档级表示时,需要以某种方式聚合词向量。文献中常见的方法是简单地取文档中所有术语的词向量的平均值,或者将向量连接成固定大小的文档向量(Yang、Macdonald 和 Ounis,2017 年)。还提出了从tf-idf加权词向量平均值得出的文档表示(Corrêa Júnior、Marinho 和 dos Santos,2017 年;Zhao、Lan 和 Tian,2015 年)。另一种方法在训练词向量的同时训练文档级密集向量表示(Le 和 Mikolov,2014 年)。我们将后一种方法称为 doc2vec。

许多研究已将神经词嵌入应用于分类和语义评估任务。例如, Billah Nagoudi、Ferrero 和 Schwab (2017)应用词嵌入来模拟阿拉伯语句子之间的语义相似性。提出了三种不同的句子级别聚合,即句子中所有单词的词向量之和、词向量的逆文档频率加权和以及词性加权和。作者发现加权和表示提供了更准确的句子相似性。在另一项研究中, Corrêa Júnior 等人。(2017)使用具有不同特征表示的分类器集合开发了一种用于情感分析的分类方法,即tf-idf矩阵、平均词向量表示和词向量的tf-idf加权平均值。最近,李等人。 (2017)在包含 3.9 亿条英文推文的 Twitter 数据集上发布了多个预训练的 word2vec 模型,并进行了一系列预处理步骤。嵌入表示在涉及 OSN 数据的 NLP 任务中的应用越来越广泛。

除了单词和文档嵌入之外,还提出了字符级嵌入模型并将其应用于 Twitter 数据,

创建tweet2vec （Dhingra、Zhou、Fitzpatrick、Muehl 和 Cohen，2016 年）。tweet2vec的动机是社交媒体数据很嘈杂，充斥着拼写错误、缩写、首字母缩略词和特殊字符，这可能导致词汇量过大。Tweet2vec将每条推文的字符序列作为输入，并将它们传递给双向 GRU 神经网络编码器以创建固定维度的推文嵌入向量。然后，将此推文嵌入传递给线性 softmax 层以预测推文的主题标签。该算法根据主题标签分类性能进行了评估。虽然这种方法可能有望创建有用的推文嵌入，但它假设主题标签是推文的有效标签。这个假设可能不成立，因为其他文本、用户提及和 URL 在定义推文主题方面也很重要，并且推文可以有多个主题标签。

最近，有人提出了对词嵌入进行语境化扩展的方法。传统词嵌入面临的一个挑战是多义性，即一个词根据上下文具有多个含义。Peters等人 (2018)提出了一种深度语境化词嵌入模型，该模型模拟了词语用法的句法和语义特征，以及这些用法在不同语言语境中的变化。该方法涉及将从双向 LSTM 训练的嵌入向量与语言模型目标相结合。该方法名为ELMo （来自语言模型的嵌入），为每个标记分配一个嵌入向量，该向量是整个输入句子的函数。这种技术可能对社交媒体文档的聚类有用。

除了迄今为止讨论的文档聚类和主题建模方法外，还开发了一系列基于深度学习的新聚类方法（Min 等人，2018 年）。其中许多技术使用深度神经网络来学习与聚类同时训练的特征表示。示例包括几个具有聚类层的深度自动编码器网络，其中损失函数是重构损失和聚类损失的组合。从文档聚类的角度来看，基于生成模型（如变分自动编码器和生成对抗网络)的聚类方法看起来很有前景，因为它们也可以从聚类中生成代表性样本。然而，迄今为止，这些技术的重点一直是图像数据集。

针对 OSN 文本数据提出了许多文档聚类和主题建模方法。这些方法通常涉及使用tf-idf矩阵或其他技术创建文档级特征表示，然后使用聚类方法将文档分组为语义相关的集群。然而，据我们所知，这些方法有很多变化，并且词嵌入表示尚未在 OSN 数据中的文档聚类任务上得到有效应用和基准测试。

3.方法

在本节中，我们描述所使用的三个数据集和处理步骤、特征表示和聚类算法以及所使用的评估措施，并讨论其属性。

应用于 OSN 数据的文档聚类和主题建模方法通常涉及几个处理步骤，如图 1 所示。首先从源中提取数据。从原始数据集或 OSN 平台 API 中提取由单个用户的文本数据组成的文档。推文和 Reddit 父评论就是文档的示例。然后处理文本元素以删除常用标点符号和停用词，并进行标记。创建每个文档的特征表示，

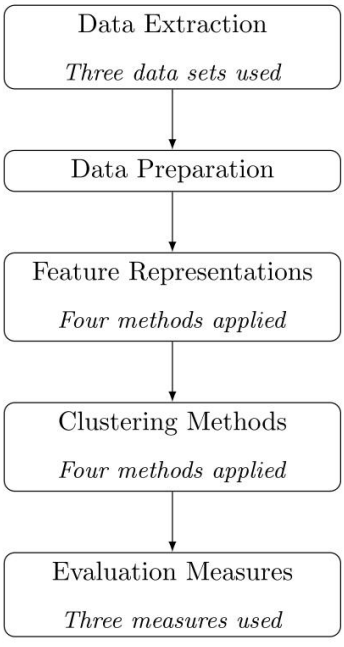

图 1.文档聚类的流程管道。本文的贡献是使用三种评估方法对三个数据集上的四种特征表示方法和四种聚类方法进行了评估。

表 1本研究
中使用的数据集、特征表示和聚类方法以及外部评估措施的概述。对于这三个数据集,我们用三种评价指标来评价特征表示和聚类方法组合以及LDA主题模型（17种组合）。

数据集
Twitter 流经#Auspol 过滤,29,283 条推文 RepLab 2013 竞赛 Twitter 数据,2657 条推文 2015 年 5 月的 Reddit 数据,40,000 条家长评论
方法
特征表示： FR1 tf-idf矩阵,其中包含每个文档的前 1000 个术语 FR2平均 word2vec 矩阵 FR3按前 1000 个tf-idf分数加权的平均 word2vec 矩阵 每个文档的FR4 doc2vec 矩阵 聚类方法： CM1 k 均值聚类 CM2 k-medoids 聚类 CM3层次凝聚聚类 CM4非负矩阵分解 (NMF) 主题模型： LDA潜在狄利克雷分配主题模型
评价措施
NMI标准化互信息 AMI调整互信息 ARI调整兰特指数

然后是聚类方法。然后使用基本事实标签计算外部聚类评估指标。表1概述了该过程每个步骤的变化。在本节的其余部分,我们将详细介绍图1 中每个步骤的方法。

3.1.数据提取

我们使用了三个OSN数据集进行评估;两个 Twitter 数据集和一个 Reddit 数据集。我们使用了 Twitter 数据,因为它已广泛用于主题建模和文档聚类的文献中。虽然使用 Reddit 数据的研究似乎较少,但 Reddit 仍然是用于主题建模和文档聚类的 OSN 数据的宝贵来源。

Reddit 也更多地被用作讨论论坛,评论的文档长度范围比 Twitter 数据的范围更广。所有三个数据集均已提供（Curiskis 等人提交）。

Twitter 数据为简短的、主题性的用户驱动内容提供了易于访问的数据源。它也广泛用于研究目的,但由于推文长度短以及使用主题标签、首字母缩略词、用户提及和 URL,因此面临许多挑战（Stieglitz 等人,2018）。第一个 Twitter 数据集是通过 Twitter 的公共 API 收集的。它是通过过滤 Twitter 流中的主题标签 #Auspol 构建的,该标签在澳大利亚经常用于政治讨论。 OSN 数据上文档聚类的一个常见应用是获取一组与特定主题相关的文档并发现主题,例如 Twitter 数据中的健康主题研究（Paul & Dredze,2014）。 #Auspol Twitter 数据集适合比较文档聚类方法,因为主题标签广泛用于链接大量不同的讨论,通常带有与澳大利亚舆论相关的其他主题标签。数据收集于 2017 年 6 月 13 日至 9 月 2 日期间,包含 1,364,326 条推文。我们通过仅选择英语推文来过滤此数据集,并根据retweeted_status字段和文本过滤器删除转发。这导致了 205,895 条推文。

该数据集不存在真实主题标签,因此我们使用一组高计数主题标签作为真实标签。我们进一步从数据集中删除了搜索主题标签 (#Auspol),因为所有推文都有此标记。一条推文上有多个主题标签是很常见的,因此为了避免主题重叠,我们删除了包含多个热门主题标签的推文。

我们还手动删除了一些相关的主题标签,例如与#marriageequality 密切相关的#ssm（同性婚姻）;我们保留了后者,因为它在更多推文中使用。最后,我们按至少有 1000 条推文的主题标签进行筛选,以保持主题相对平衡。结果为 29,283 条推文,其中 13 个主题标签表示主题标签,如表2 所示。

第二个 Twitter 数据集取自 RepLab 2013 竞赛（Amigó 等人,2013 年）。该竞赛的重点是监控实体（公司和个人）的声誉,涉及命名实体识别、极性分类和主题检测等任务。本次竞赛中使用的推文由几位受过训练的注释者在声誉专家的监督和监控下标注了主题标签。就本文而言,这些推文中标注的主题被视为黄金标准。我们使用这个数据集是因为它已经标注了黄金标准标签,并且已用于主题检测任务。

我们从 RepLab 2013 竞赛提供的主题检测任务的训练和测试数据集中下载了 Twitter 标识符列表,并于 2019 年 1 月 19 日通过 Twitter API 检索了详细信息。在 110,344 条已发布的带有标签主题的推文标识符中,我们只能检索 23,684 条推文的推文文本和其他信息。这是

表2
#Auspol Twitter 数据集中每个主题标签的推文数量。

主题编号	井号	推文
1	#qldpol	3845
2	#qanda	3592
3	#insiders	3495
4	#lnp	3434
5	#politas #婚	2618
6	姻平等 #springst #nbn	2562
7	#trump	1708
8	#uspoli	1626
9	#stopadani	第1547章
10		1498
11	#climatechange	1186
12	#turnbull	1148
十三		1024

可能是因为推文和用户在发布后被删除。此外,这个标签中还有一长串主题
数据。事实上,对于 23,684 条推文,总共有 3432 个不同的主题,其中 1263 个主题包含一条推文。为了确保
由于有足够的数据点供我们的方法检测,我们将每个主题的频率计数限制为 100。我们还删除了
标签表示“其他主题”,因为这并不代表内部一致的主题。经过此过滤后,我们得到了 2657 个数据集
带有 13 个主题标签的竞赛推文。表3给出了所使用的主题标签列表。

我们最初纳入 RepLab 2013 数据集主要是因为主题发现的比较结果可从
竞赛。然而,由于Twitter的API无法检索大量推文,因此无法进行准确的比较
不再可能了。尽管如此,真实主题标签仍然允许对方法的性能进行基准测试。

第三个数据集来自 Reddit 平台,由 2015 年 5 月以来 Reddit 子 Reddit 页面的家长评论及其相关评论组成。Reddit 平台广泛用于与特定主题或主题相关的讨论,按以下分组:
subreddit 页面,因此非常适合本研究。此外,Reddit 评论可能比推文更长。Reddit 父评论指的是
热门评论,可能有其他用户回复,也可能没有。该数据已在 Reddit 网站上公开
（Reddit,2015 年）。完整数据集包含 50,138 个 subreddit 页面上的约 5450 万条评论。我们之所以选择这个数据集,是因为它
全文免费提供,包含多个主题的讨论。因此,它是用于基准测试方法的理想数据集。
我们选择了五个代表不相交主题的 Reddit 子页面进行分析。这五个 Reddit 子页面也曾在之前的文章中使用过
研究基准分类模型（Gutman & Nam,2015）。由于家长的评论和回应本质上是相关的,我们
将所有用户帖子汇集到按父评论标识符分组的文档中。表 4显示了每个家长的评论数
subreddit 页面。我们从五个 subreddit 页面中随机抽样了 40,000 个父评论标识符,然后使用这些页面
来表示基本事实标签。

Reddit 数据在本研究中特别有用,因为它包含的每个文档的字符长度范围比 Twitter 数据更广,
因为 Twitter 对字符数有限制。通过
文档长度可以为未来研究特定数据集的最佳方法提供指导。为了检查这种性能,我们根据每个文档的字符数将 Reddit 数据划分为四个不同的子集。四人详细信息

数据分区如表5所示。为了与 Twitter 数据集进行比较,一条推文最多有 240 个字符。对于
#Auspol Twitter 数据,平均字符长度为 117,第 25 个百分位为 103,第 75 个百分位为 138。大多数推文
因此属于长度为 101 到 200 个字符的文档组。

表3
RepLab 2013 Twitter 数据集中每个主题标签的推文数量。

主题编号	话题	推文
1	出售	第329章
2	铃木杯	296
3	用户评论	262
4	洗钱/恐怖主义融资	199
5	YouTube 观看次数记录	195
6	狂热粉丝 - Beliebers	154
7	普林斯顿进攻	131
8	出售 - 日产汽车,配件分析	127
9	笑话	127
10	体育赞助商	127
11	垃圾邮件	114
12	讽刺性的批评	111
13	MotoGP - 用户评论	103

表4
每个 subreddit 页面的父评论数。

主题编号	Reddit子版块页面	家长评价
1	[redacted]	10,563
2	新闻	9488
3	pcmasterrace	9186
4	电影	6263
5	关系	4500

表5
Reddit 数据根据文档字符长度分为四组。文档按父评论分组。平均字符
给出了每个文档的长度和平均标记数。

字符长度范围	文件数量	平均字符长度	平均代币数量
1–100	15,273	46.1	4.5
101–200	8360	144.9	13.3
201–500	9310	317.4	28.6
501 或以上	7057	1,584.5	141.1

3.2 数据准备

本研究中的数据准备和分析使用Python 3.6.1进行。对于文本预处理,我们删除了
nltk 3.2.4包中的停用词和字符串中的标点符号。为推文创建了一个定制的分词器函数，
保留标签和用户提及,并删除 URL。为了标记 Reddit 数据,我们只需删除标点符号和标准
停用词。我们没有应用任何词干提取或词形还原。我们还使用了sklearn 0.19.1中的TfidfVectorizer函数来实现tf-idf方法和加权 word2vec 方法。

对于 #Auspol Twitter 数据,我们从文本中删除了 14 个作为真实标签的标签列表,此外
#Auspol Twitter API 搜索查询。 RepLab 2013 Twitter 数据集带有注释的主题标签,这些标签不直接基于任何
单个 token,因此无需修改。对于 Reddit 数据,由于 subreddit 页面被用作基本事实标签,因此我们
不需要修改文本。

3.3. 特征表示

在本研究中,我们评估了四种结合构建文档特征表示的方法的性能
四种常用的聚类算法。我们还将 LDA 主题模型纳入单独的主题模型类别,因为
该技术仅将词袋矩阵作为输入。这些方法概述在表1 中,其中每个方法组件都有一个
代码方便参考。四种特征表示被编码为FR1-FR4 ,四种聚类方法被编码为CM1-
CM4和 LDA 主题模型被简单地编码为LDA。虽然文献中提出了许多其他技术,例如
模因识别研究（JafariAsbagh 等人,2014 年;Shabunina & Pasi,2018 年）,我们没有实施它们进行评估,因为它们
特定于 Twitter 的数据。不过,我们在讨论中提供了比较结果,这些结果来自其他
学习。

对于FR1， tf-idf矩阵被限制为按频率计算的每个文档的前 1000 个词条,因为没有性能改进
通过包含更多术语而获得。这可能是由于社交媒体文本的简短性质,它产生稀疏的tf-idf特征向量；
频率较低的术语通常对聚类没有用。

word2vec 模型是一种神经网络,经过训练可以为语料库中的每个标记创建具有固定维度的密集向量。
预先训练好的 word2vec 模型可用于 Twitter 数据（Godin、Vandersmissen、De Neve 和 Van de Walle,2015）,我们发现它
在本研究中使用的 Twitter 数据集上表现不佳。一个问题是数据中的许多标记不在经过训练的范围内
模型的词汇量,以及单词之间的语义关系在不同的数据集上可能会有很大的不同。此外，
没有针对大量 Reddit 数据的预训练模型。此外,这些中还有很多超参数
因此,为不同的数据集找到一组理想的值是一个有用的贡献。由于这些原因,我们训练了自己的单词
嵌入和文档嵌入模型。

FR2和FR3中使用的 word2vec 模型采用连续词袋 (CBOW) 方法进行训练（Mikolov 等人，
2013）， 100 个维度,大小为 5 的上下文窗口,最小字数为 1。我们测试了这些超参数的变体，
包括从 3 到 15 的上下文窗口大小、更高的维度和最小字数。我们发现变化在
使用三种聚类评估措施的性能是最小的,并且选择的超参数是最优的。一些
考虑到社交媒体文本的文档长度较短,这些结果是有意义的。我们得出结论,word2vec 的 100 个维度是
足以代表短文档的单词。每条推文的平均标记数为 9,第 75 个百分位为 11,因此
大小为 5 的上下文窗口捕获了大多数推文的所有标记。然而,我们确实发现训练次数存在显着差异
三个数据集使用的时期。我们将在第 4.1 节中报告此分析。对于所有其他超参数,我们使用默认

gensim 3.4.0 python 包提供的值（ eh ek & Sojka,2010）。

FR2是通过对每个文档中每个标记的词向量取元素平均值来构建的,返回一个 100 维的密集特征向量。FR3是通过对文档中每个单词的词向量取tf-idf加权平均值来构建的。使用的tf-idf矩阵是FR1中构建的按频率排列的前 1000 个词矩阵。此过程排除了不在前 1000 个tf-idf词中的任何词向量,尽管再次尝试了更多数量的顶级词,但发现使用的评估指标有所下降。我们将在第 3.5 节中讨论使用的评估指标。

doc2vec 模型是一种神经网络,经过训练可以为语料库中的每个文档创建一个具有固定维度的密集向量。FR4 中的 doc2vec 模型使用分布式词袋法(dbow) 进行训练,维度为 100,上下文窗口大小为 5,最小字数为 1。之所以使用分布式词袋法,是因为它可以在同一个嵌入空间中训练词向量和文档向量(Le & Mikolov, 2014),这对于解释文档嵌入非常有用。与 word2vec 模型一样,我们测试了超参数的变化,发现评估指标因训练周期数而存在显著差异,不同的数据集具有不同的最佳周期数。这与Lau 和 Baldwin (2016)的结果类似,其中对 430 万个单词进行训练的dbow doc2vec 模型的最佳周期数为 20,而对于大小为 50 万个单词的数据集,最佳周期数为 400。 Lau 和 Baldwin (2016)还发现最佳维数为 300,窗口大小为 15。我们方法的最佳值较低可能是因为 OSN 数据的文档长度较短,以及我们的数据集（尤其是 Twitter 数据)的字数较少。

3.4.聚类方法

对于聚类方法,我们选择了文献中常用的四种技术（Klinczak & Kaestner,2016;Naik 等人,2015）,它们也在我们的数据集上给出了可比较的结果。首先,我们应用了使用欧几里德度量和最多 100 次迭代的k 均值聚类算法(CM1) 。该算法使用不同的随机种子对数据运行多次。

CM2指的是 k-medoids 算法。为此,我们使用了pyclustering 0.8.2 python 包,其中的起始质心根据均匀分布进行采样。Klinczak和 Kaestner (2016) 使用了 k-means 和 k-medoids 聚类。对于CM3,我们应用了具有欧几里得度量和 Ward 链接的层次凝聚聚类算法。Ferrara等人 (2013)使用层次凝聚聚类来聚类相似性矩阵。对于CM4,我们使用了非负矩阵分解 (NMF) 算法,为此我们使用了sklearn 0.19.1包中的默认参数。NMF 已在 OSN 数据中的主题建模中得到多种应用（Godfrey 等人,2014;Klein 等人,2018）。对于聚类方法和 LDA 模型,我们将聚类或组件的数量设置为等于评估数据中唯一标签的数量。与Klinczak 和 Kaestner (2016) 一致,我们用一系列超参数测试了 DBSCAN 聚类算法,但发现它对所有特征表示的性能都很差。文档要么被分组为异常聚类,要么被分组为大量非常小的聚类。一个可能的原因是特征表示是高维且稀疏的,因此使用基于密度的方法可能无法很好地聚类。

LDA主题模型经过 10 次训练,块大小为 10,000,并更新每条记录。我们再次使用gensim 3.4.0包中其他超参数的默认值。我们包含此方法,因为它常用于文档聚类和主题建模。为了给每个文档分配一个主题标签,我们选择了概率最高的主题。

3.5.评价措施

用于评估文档聚类方法的指标通常分为两类:内在指标和外在指标。
内在指标（例如聚类分离和凝聚度指标)不需要真实标签。此类指标描述了聚类内和聚类之间的差异。但是,它们依赖于所使用的特征表示,因此对于使用不同特征集的方法,它们无法提供可比较的结果。外在指标需要真实标签,但可以跨方法进行比较。常见的外在指标包括准确率、召回率和 F1 （Naik et al., 2015）,但这些指标依赖于聚类标签与真实标签的排序,而这在大量标签的情况下是一个问题。在这种情况下,诸如互信息和兰德指数之类的指标更为合适,因为它们与标签的绝对值无关。

互信息是两个离散随机变量之间相互依赖性的度量。它量化了在了解另一个离散随机变量的情况下,对一个离散随机变量的不确定性的减少。互信息量高表示不确定性大幅减少。对于具有联合概率分布p(x, y) 的两个离散随机变量X和Y ,互信息MI(X, Y) 定义为

$$MI(X,Y) = \sum_{y\in Y}\sum_{x\in X} \log\left(\frac{p(x,y)}{p(x)p(y)}\right).$$

常用的测量方法是标准化互信息 (NMI),它将 MI 标准化为 0 到 1 之间的值,其中 0 表示没有互信息,1 表示一致。这对于比较不同方法和研究的结果很有用。NMI 如下所示。

$$NMI(X,Y) = \frac{\text{混合函数}(H(X),H(Y))}{\sqrt{\text{华夏华}(H(X)H(Y))}},$$

其中H(X)和H(Y)表示边际熵,由下式给出

$$H(X) = -\sum_{i=1}^{n} P(i)\log(P(i)).$$

兰德指数是标签和簇之间相似性的配对计数度量。它也取 0 到 1 之间的值，
其中 0 表示随机标记，1 表示相同标记。给定一组元素 $S = \{s_1, \ldots, s_n\}$，其中S需要比较，$X = \{X_1, \ldots, X_r\}$ 和 $Y = \{Y_1, \ldots, Y_s\}$ 和两个分区
Rand 指数表示分区X和Y出现的频率

$$RI(X, Y) = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}},$$

其中a表示S中位于X中相同子集和Y中相同子集的元素对的数量， b表示
S中位于X的不同子集和 Y 的不同子集的元素对的数量。a和b的值一起给出了数量
分区一致的次数。值c表示S中位于X的同一子集中的元素对的数量
和 Y 的不同子集， d给出S中位于X的不同子集和Y的同一子集的元素对的数量。

为了使外在聚类评估措施可用于跨方法和研究的比较，此类措施需要一个固定的
边界和恒定的基线值。NMI 和 RI 均被缩放为具有 0 到 1 之间的值，因此满足第一个条件。
然而，事实证明，这两个指标都会随着标签数量的增加而单调增加，即使是任意聚类
分配（Vinh、Epps 和 Bailey，2010）。这是因为互信息和兰德指数都没有一个常数
基线，这意味着这些度量在具有不同数量的聚类方法之间是不可比较的。为了考虑
为此，提出了 MI 和 RI 的调整版本。调整后的兰特指数 ARI 通过其预期值调整 RI：

$$ARI(X, Y) = \frac{RI(X, Y) - E\{RI(X, Y)\}}{\max\{RI(X, Y)\} - E\{RI(X, Y)\}}$$

其中E{RI(X, Y)} 表示RI(X, Y)的期望值。 ARI 取值在 0 到 1 之间，其中 1 代表相同
分区，并根据X和Y中的分区数进行调整。以类似的方式，调整后的互信息 AMI 由下式给出：

$$AMI(X, Y) = \frac{MI(X, Y) - E\{MI(X, Y)\}}{\max\{MI(X, Y)\} - E\{MI(X, Y)\}},$$

其中E(MI(X, Y)) 表示 MI 的预期值（Vinh 等人，2010）。 AMI 取 0 到 1 之间的值，其中 1
代表相同的分区，并根据使用的分区数量进行调整。确保可比性的最佳措施
评估的主要指标是 AMI 和 ARI。下一个问题是这两个指标如何相互比较。通过开发
关于广义信息理论测度的理论， Romano、Vinh、Bailey 和 Verspoor (2016)得出结论，AMI 是
当标签不平衡并且存在小簇时，这是更好的措施，而当标签不平衡时，应该使用ARI
具有大且相似大小的体积。

在本文中，我们报告了 AMI、ARI 和 NMI 措施。之前的许多研究都报告了 NMI 测量，因此对于
出于比较目的，我们将其纳入我们的评估中。鉴于本研究的数据和方法，ARI 可能更准确
那么 AMI 就很合适，因为表 2和表 4表明，文档在标签之间的分布相对均衡。我们仍然
包括 AMI，因为看看结果与 NMI 有多大差异是很有趣的。

由于本研究中使用的数据集较短且噪声较大，我们研究了不同随机种子对
表现。我们使用不同的随机种子运行每种方法 20 次，计算 NMI、AMI 和 ARI 的平均值，并绘制
这些措施的分布。

4.结果

在本节中，我们将介绍我们的分析结果。我们首先描述了最佳时期数的结果
word2vec 和 doc2vec 嵌入表示，应用于所有三个数据集。然后我们评估所有
方法。最后，我们讨论使用 doc2vec 特征表示来解释主题的方法。

4.1.嵌入模型的最佳训练时期

训练神经网络模型的一个关键超参数是时期数。 epoch 太多，模型可能会过度拟合
数据太少，性能可能会很差。我们首先探索了平均 word2vec 模型（FR2和
FR3)和 doc2vec 模型(FR4)的 epoch 数。这些结果为研究真实主题提供了指导
标签不存在。我们使用 k 均值聚类(CM1)作为聚类方法，因为它为嵌入提供了最佳结果
交涉。对于 25 到 300 之间的每个纪元值，增量为 25，我们使用不同的方法训练模型 20 次
随机种子并根据地面真实标签进行评估。这是对所有三个数据集进行的。表 6总结了最佳方案
按方法和数据集划分的纪元结果。对 #Auspol Twitter 数据的分析图如图2 (a) 和
RepLab 2013 数据如图2(b) 所示。Reddit 数据的结果如图3所示。为了节省空间，我们仅评估了 AMI 和
ARI 对 Reddit 数据进行测量。这是因为 AMI 通常会给出与 NMI 类似的结果，但会进行机会调整。

对于图2（a)中的 #Auspol 数据，很明显 doc2vec 给出了最佳结果，并且在 75 个 epoch 左右达到性能峰值。

表6

word2vec 和 doc2vec 方法在三个数据集上的最佳训练周期数。

数据集	doc2vec	wtd.word2vec	unwtd.word2vec
推特#Auspol	75	250	250
Twitter 代表实验室 2013	300	200	200
Reddit:1-101	175	75	50
红迪网:101–200		100	200
Reddit:201-500	150100	50	50
Reddit:501 +	50	25	二十五

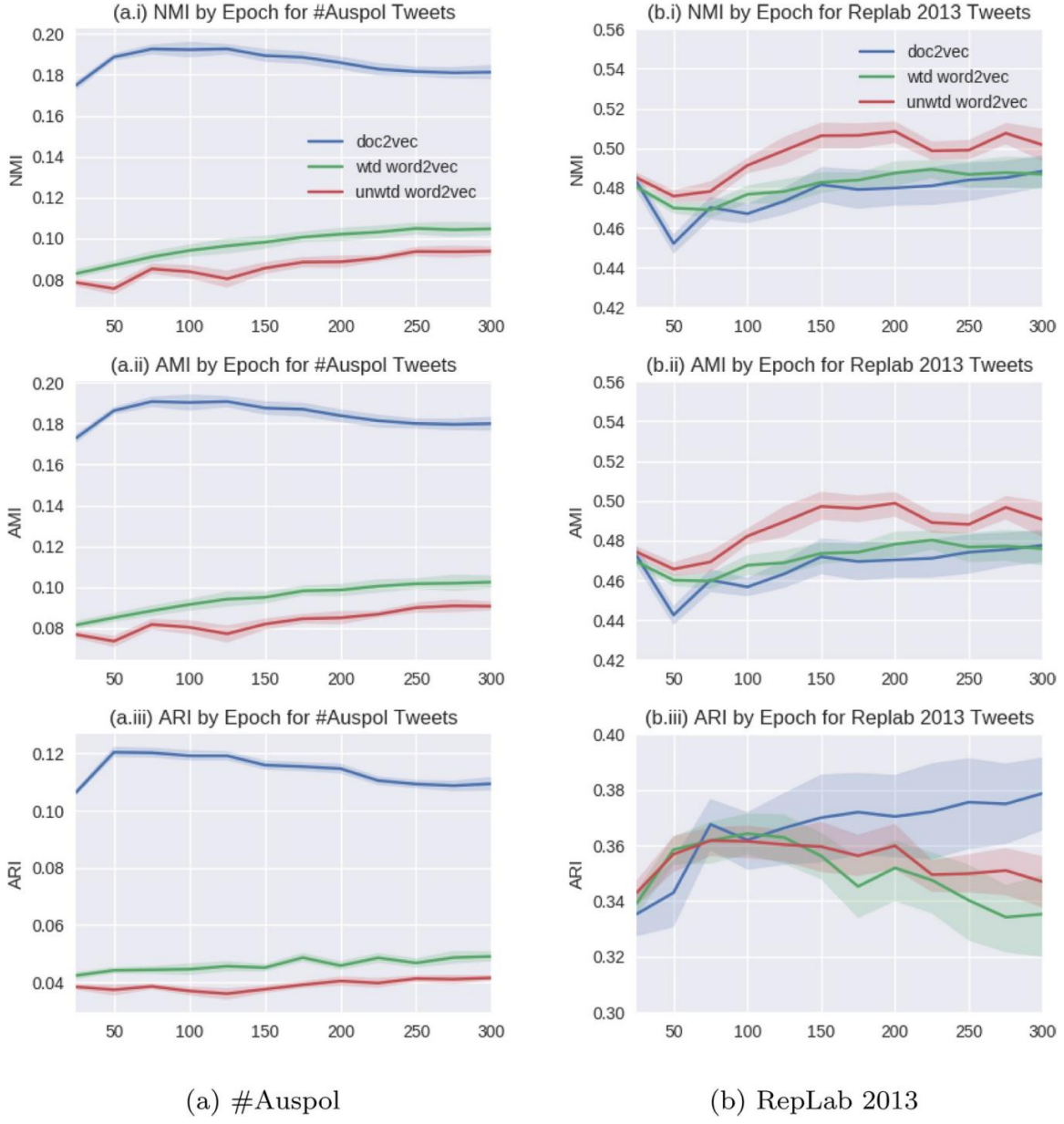

图 2.使用 k 均值聚类在 Twitter 数据上运行 20 次 word2vec 和 doc2vec 表示,通过训练纪元（水平轴)绘制三个评估指标（垂直轴）。 (a) 显示 #Auspol Twitter 数据上的结果,(b) 显示 RepLab 上的结果
2013 年推特数据。显示了基于不同随机种子的 95% 置信带。

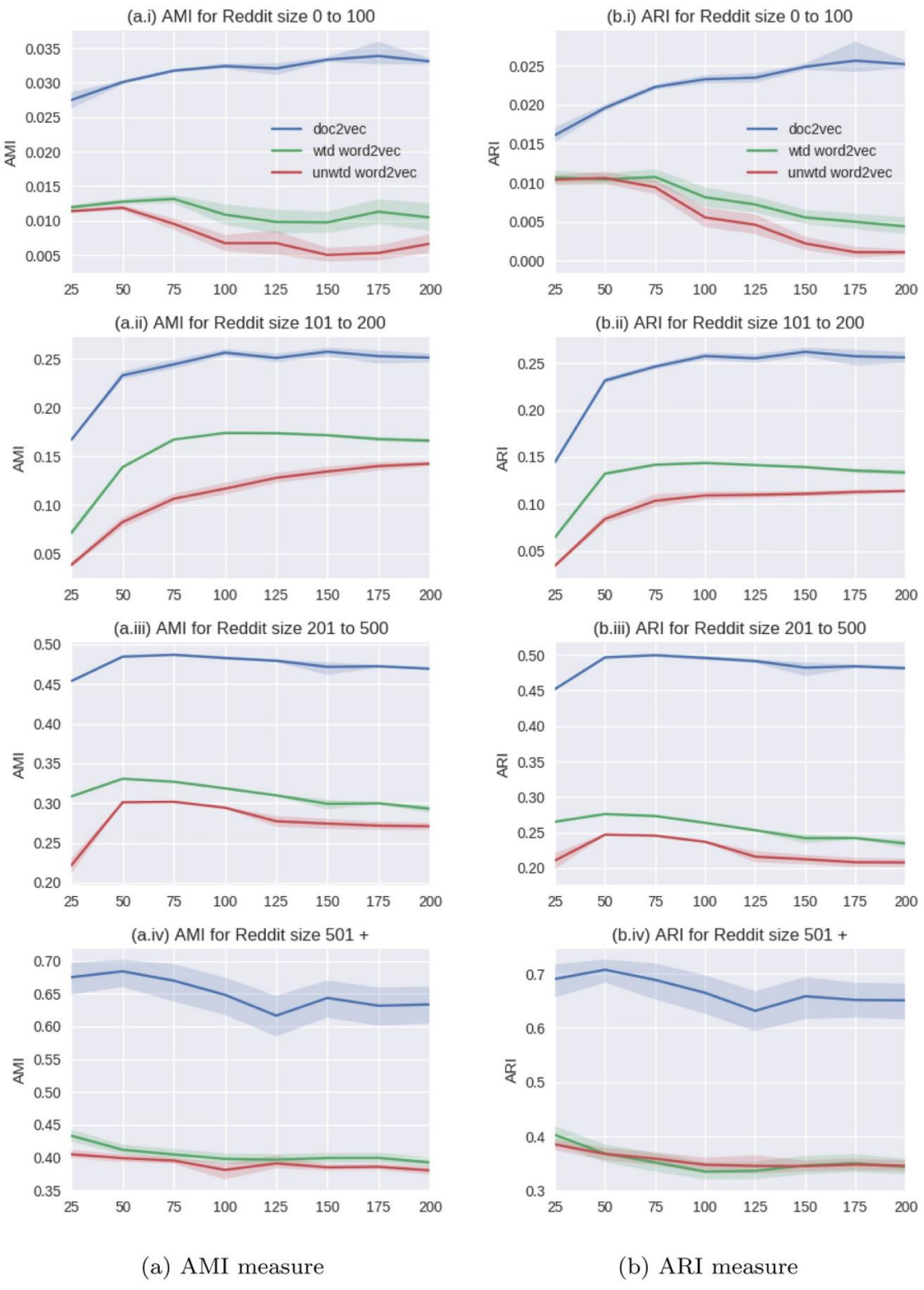

图 3.使用 k-means 聚类对 Reddit 数据集上的 word2vec 和 doc2vec 表示进行 20 次运行，通过训练时期 (横轴) 绘制 AMI 和 ARI 评估指标 (纵轴) 图。行中给出了按大小范围划分的不同 Reddit 数据集。列 (a) 显示 AMI 结果，(b) 显示 ARI 结果。显示了基于不同随机种子的 95% 置信区间。

word2vec 方法通常可以通过更多 epoch 提供更好的性能,最大值约为 250。 tf-idf
加权平均word2vec方法比非加权平均word2vec方法表现更好,并且其性能提高了
比无加权方法更平滑。由于 95% 置信区间较窄,因此种子之间的差异也不大。

对于图2(b)中的 RepLab 2013 数据,结果有很大不同。未加权平均 word2vec 方法给出了最好的
NMI 和 AMI 指标的表现。然而,在 ARI 测量中,两种 word2vec 方法的性能均下降
100 个 epoch 后,doc2vec 方法有所改进。这可能是由于 word2vec 模型对数据的过度拟合造成的,
这很可能是因为 RepLab 2013 的数据比 #Auspol 的数据小得多。 ARI 指标也是首选指标
标签数量较大且平衡（Romano 等人,2016 年）。该数据集相对平衡（见表3），
因此,ARI 是比 NMI 和 AMI 更合适的绩效衡量标准。总体而言,根据 RepLab 2013 数据,最佳
word2vec 方法的 epoch 数量为 200,而 doc2vec 方法的最佳值为 300。
doc2vec 方法的最佳时期并不奇怪,因为它也在训练文档向量,因此具有比
word2vec。

转向图3中四个 Reddit 数据集的结果,doc2vec 方法再次给出了最佳性能。此外,
doc2vec 有一个明显的模式,较短的文档需要更多的训练周期才能达到最佳性能。为了
对于少于 100 个字符的文档,使用 k-means 聚类的 doc2vec 的性能提高了约 250 个 epoch。
对于包含 101-200 个字符的文档,这个数字下降到 150 个 epoch,对于文档长度较大的数据集,这个数字下降到 150、100 和 50。
增加订单。这一观察到的模式与Lau 和 Baldwin (2016)的结果一致,证实 doc2vec 模型需要更少的
在较大的文档上训练 epoch。

对于 word2vec 方法， tf-idf加权平均词向量方法比非加权平均方法效果更好
方法。这与之前研究的结果一致（Billah Nagoudi 等人,2017 年）。在最短文档范围内,两种方法
通过更多的训练,性能几乎没有提高,但加权后的 75 个 epoch 中,这两项指标都下降了
word2vec 方法和 50 为未加权方法。对于这种下降的一种可能的解释是,平均词向量可能只
超越言语的界限才有意义。对于此大小范围,每个文档的平均字数为 4.5,这可能太高了。
低。在长度为 101 到 200 个字符的文档中,加权 word2vec 方法表现更好,但所需的
训练时期。这些结果也与 Twitter 数据集上的结果类似,后者通常具有相似的字符长度
范围。在最大的文档上,两种方法都需要 25 个或更少的时期才能达到最佳性能。

通过此分析,很明显,doc2vec 方法始终比平均 word2vec 方法提供更好的性能
方法,除非数据集的文档数量较少。此外,训练时期的数量
doc2vec 通常与文档大小成反比,需要更多 epoch 才能达到最佳性能
文档大小较小。Doc2vec 也需要比 word2vec 更多的训练周期。然而,这些关系并不
观察 #Auspol Twitter 数据,其中 doc2vec 最佳 epoch 数为 75,低于 word2vec 最佳 epoch 数 200。
RepLab 2013 数据上 doc2vec epoch 的最佳数量要高得多,为 300。一个解释可能是,虽然
doc2vec 模型通过在 #Auspol 数据上进行更多训练,改进了其内部损失函数,但这些改进并没有
导致聚类任务上的表现更好。这可能是因为使用的标签可能有一些重叠
贡献项。对于 word2vec 方法,通常通过tf-idf分数加权可以提高性能,并且需要更少的
训练纪元。然而,考虑到最短 Reddit 文档的低峰值,应注意 epoch 的数量。

4.2.使用聚类措施进行绩效评估

在本节中,我们提供了四种聚类方法对四种特征表示的平均评价指标,以及

表7
使用归一化互信息对 #Auspol Twitter 数据的特征表示和聚类方法进行性能评估
(NMI)、调整互信息 (AMI) 和调整兰特指数 (ARI) 指标。

特征表达	聚类	[illegible]	[illegible]	阿里
doc2vec	分层的	0.165	0.154	0.059
	k 均值 k 中	0.193	0.191	0.120
	心点	0.107	0.105	0.064
	纳米纤维	0.102	0.100	0.056
wtd word2vec	分层 k 均值 k 中	0.088	0.079	0.021
	心点	0.105	0.102	0.047
		0.043	0.016	0.001
	纳米纤维	0.062	0.058	0.030
unwtd word2vec	层次结构	0.085	0.076	0.020
	k-均值	0.094	0.090	0.041
	k-中心点	0.043	0.019	0.001
	[illegible]	0.058	0.054	0.025
TF-IDF	分层 k 均值	0.163	0.085	0.013
		0.114	0.070	0.014
	k-中心点	0.079	0.028	0.004
	[illegible]	0.132	0.110	0.032
[illegible]	LDA	0.043	0.041	0.021

表8
使用 NMI、AMI 和 ARI 度量对 RepLab 2013 Twitter 数据上的特征表示和聚类方法进行性能评估。

特征表示	聚类	[illegible]	[illegible]	阿里
doc2vec	分层 k 均值	0.449	0.437	0.313
		0.488	0.478	0.379
	k-中心点	0.290	0.278	0.215
	纳米纤维	0.261	0.249	0.152
wtd word2vec	分层 k 均值 k 中	0.506	0.491	0.330
	心点	0.488	0.478	0.352
		0.421	0.404	0.274
	纳米纤维	0.401	0.384	0.266
unwtd word2vec	分层的	0.519	0.507	0.347
	k-均值	0.508	0.499	0.360
	k-中心点	0.435	0.414	0.278
	纳米纤维	0.425	0.407	0.286
TF-IDF	分层 k 均值 k 中	0.466	0.417	0.203
	心点	0.450	0.379	0.179
		0.192	0.075	0.011
	纳米纤维	0.437	0.427	0.348
[illegible]	LDA	0.180	0.169	0.140

LDA 模型,每个方法在每个数据集上有 20 个不同的种子。我们还包括分布图来说明变异性
在性能上。

表 7提供了 #Auspol Twitter 数据集上每种方法的三个评估指标的平均值。我们设定
doc2vec 方法的最佳 epoch 数为 75,word2vec 方法的最佳 epoch 数为 250。从这个表中可以清楚地看出
doc2vec 特征表示与 k-means 聚类在所有三个评估指标上都优于其他方法,尤其是在 ARI 上。层次聚类对 NMI 和 AMI 的得分接近,但 ARI 得分低得多。对于 doc2vec 和 word2vec
特征表示,NMF 表现不佳。k-medoids 聚类的性能与 NMF 相似。对于 word2vec
表示,k均值聚类也给出了最佳性能。

一个有趣的观察结果是,某些方法在 NMI 和 AMI 测量之间的得分下降相对较大,
说明AMI的机会调整很重要。 tf -idf表示受此影响最大。例如,
具有分层聚类的tf-idf矩阵的 NMI 高达 0.163,远远领先于 word2vec 方法,但 AMI 为 0.085。
相比之下,doc2vec 和 word2vec 方法的下降较小。如前所述,AMI 和 ARI 更合适
由于 ARI 会根据机会进行调整,因此它比 NMI 更适合评估指标。在这个数据集上,ARI 更合适,因为
每个主题标签的推文相对相似。因此,使用 k 均值聚类的 doc2vec 表示法远远优于
其他方法。

表 8显示了使用 doc2vec 模型训练 300 个 epoch 后 RepLab 2013 Twitter 数据集的平均结果,以及
word2vec 方法经过 200 个 epoch 的训练。总体而言,其性能远高于 #Auspol 数据,这可以通过以下方式解释
RepLab 2013 数据具有更清晰的专业注释主题。在ARI评分上,带有k-means的doc2vec方法
聚类表现最佳,但采用层次聚类的无加权 word2vec 方法对 NMI 的表现更佳
和 AMI 测量。对此的一个解释是,这个数据集的大小太小,不足以让嵌入表示
训练结果准确,因此进一步训练并不一定能提高聚类性能。这反映在
图2(bi) 和 (b.ii)中明显可见下降。

为了检查平均测量值的变异性,我们绘制了特征表示方法的分布,其中
表现最佳的聚类算法和 LDA 主题模型。图 4显示了三个评估指标在
#Auspol (a) 和 RepLab 2013 (b) Twitter 数据集。在图4(a)中,采用k-means聚类的doc2vec方法明显领先
其他方法对所有三项措施的影响。两种 word2vec 方法的结果之间也存在显着重叠,
表明分数接近时需要进行多次运行。请注意,具有层次聚类的tf-idf方法没有
由于两种算法都是确定性的,因此每次运行都有相同的结果,因此在绘图上显示。

对于图4(b)中的 RepLab 2013 数据集,word2vec 方法再次显示出与 doc2vec 方法显着的重叠
表现在较低的范围内。值得注意的是,doc2vec 方法显示出两个接近的峰值。这些峰值最
对于 NMI 和 AMI 测量很重要,但对于 ARI 也存在。这可能表明 doc2vec 方法针对本地进行了优化
训练过程中,由于存在较大的差距,导致一些随机种子的运行效果不佳。
与 #Auspol 数据上的 word2vec 方法相比,doc2vec 具有更高的性能,并且两者之间的性能接近
在 RepLab 2013 上使用 word2vec 和 doc2vec,word2vec 方法比 doc2vec 更好地处理较小的 RepLab 2013 数据集。这
可能是因为 RepLab 2013 数据集中没有足够的数据点来最佳地训练 doc2vec 表示。
尽管如此,doc2vec 仍然在两个 Twitter 数据集的 ARI 测量中给出了最佳性能。

最后,我们提供了对 Reddit 数据运行这些方法的结果。图 5显示了 NMI (a)、AMI (b) 和 ARI (c) 值
Reddit 数据集上的方法。横轴比较文档长度数据分区的结果。只有最好的
显示每个特征表示的执行聚类方法。各评价指标的平均分
表 9中给出了文档长度范围 1-100 和 101-200 的方法,表 10中给出了文档长度范围 201-500 的方法
以及 501 或更高。从这些图和平均结果可以清楚地看出,doc2vec 方法在所有四个方面都提供了最佳性能

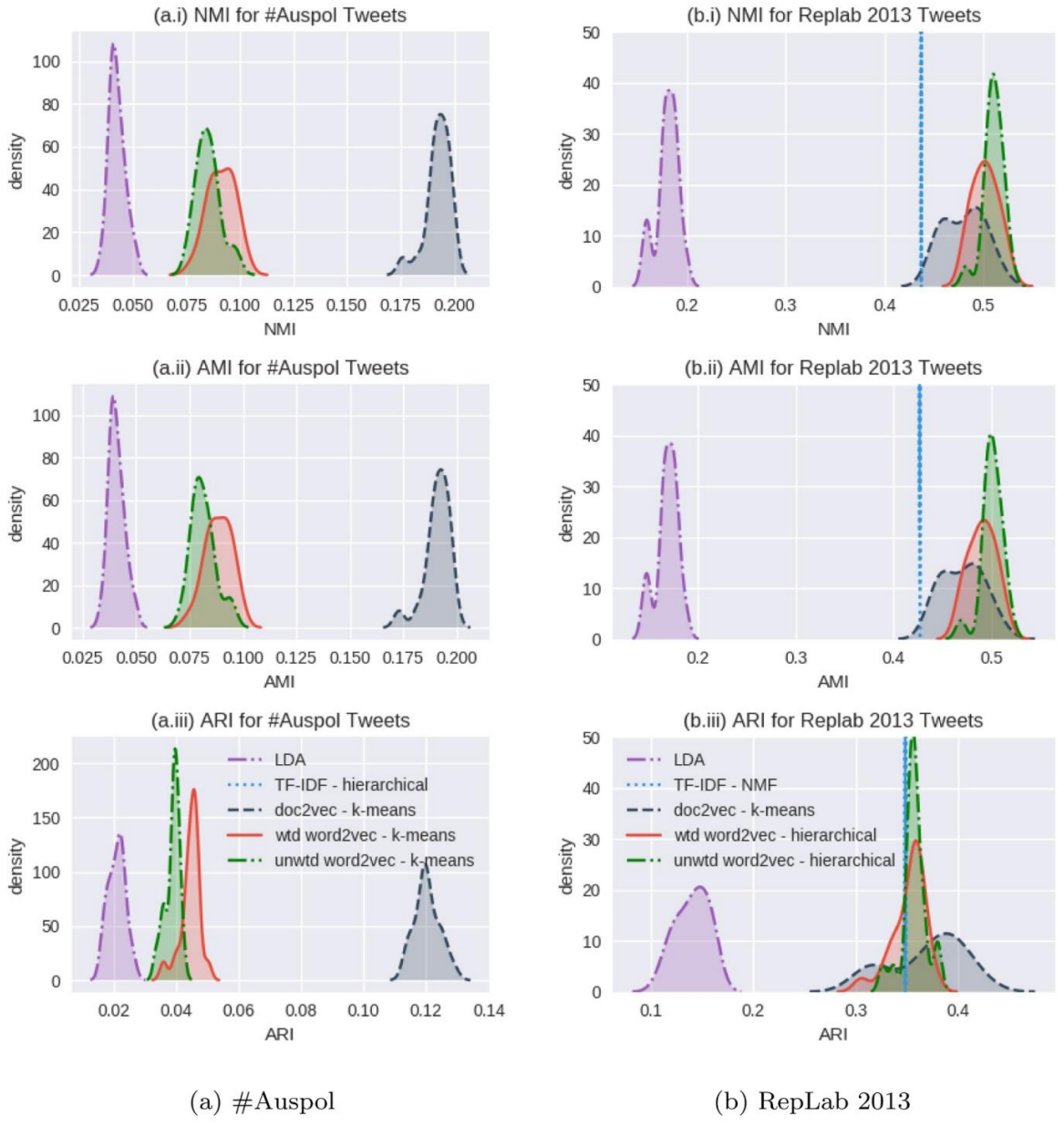

图 4.使用最佳性能聚类算法的四种特征表示的随机种子上的三个评估度量（水平轴)的密度图,并使用 LDA 进行比较。 (a) 显示了 #Auspol Twitter 数据的结果,(b) 显示了 RepLab 2013 Twitter 数据的结果。

按大小范围对数据集进行了分类。这一发现证实了 #Auspol Twitter 数据集的结果。与未加权平均 word2vec 方法相比， tf-idf加权平均 word2vec 方法始终能够提高性能。有趣的是， tf-idf方法和 LDA 模型仅在最后一个大小范围内（字符数大于 500)提供与 word2vec 方法相当的性能。

4.3.话题解读

显然,基于 ARI 测量,带有 k 均值聚类的 doc2vec 模型在 #Auspol Twitter 数据集和 Reddit 数据集以及 RepLab 2013 Twitter 数据集上表现出色。但是,主题发现模型的实用性取决于生成的主题的可解释性。在本节中,我们旨在通过更深入地分析使用 k 均值聚类的 doc2vec 表示生成的聚类来解决这个问题。

我们首先考虑 #Auspol 数据的结果,其中我们分析了文档簇与

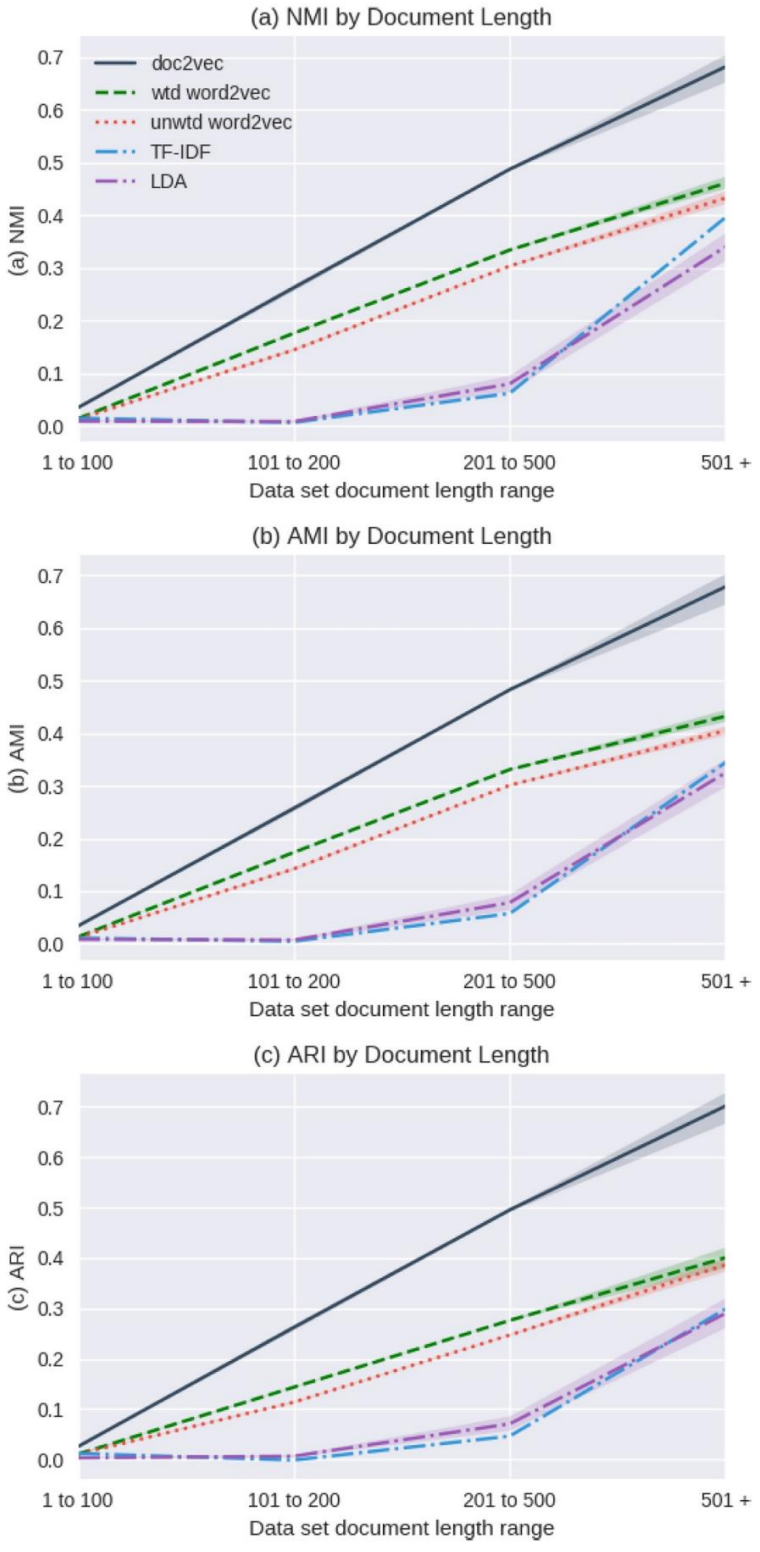

图 5.在具有不同字符文档长度的 Reddit 数据上采用最佳聚类方法的随机种子的三种评估措施的绘图。 (a) 绘制 NMI，(b) 绘制 AMI，(c) 绘制 ARI。

表9
对文档长度范围为 1-100 和 101-200 个字符的 Reddit 数据上的每种方法的性能评估。

文档字符长度	特征表达	聚类	[illegible]	[illegible]	[illegible]
1–100	doc2vec	分层 k 均值	0.029	0.027	0.017
			0.034	0.034	0.026
		k-中心点	0.012	0.011	0.004
		[illegible]	0.030	0.023	0.015
	wtd word2vec	分层 k 均值 k 中	0.013	0.012	0.010
		心点	0.014	0.013	0.011
			0.007	0.003	0.000
		[illegible]	0.010	0.009	0.000
	unwtd word2vec	层次结构	0.011	0.011	0.011
		k-均值	0.012	0.012	0.011
		k-中心点	0.007	0.006	0.000
		[illegible]	0.012	0.011	0.010
	TF-IDF	分层 k 均值 k 中	0.009	0.003	0.000
		心点	0.005	0.002	0.000
			0.005	0.001	0.000
		纳米纤维	0.014	0.011	0.012
	LDA	LDA	0.009	0.009	0.003
101–200	doc2vec	层次结构	0.115	0.111	0.067
		K 均值	0.262	0.257	0.262
		k-中心点	0.018	0.006	0.001
		[illegible]	0.127	0.096	0.027
	wtd word2vec	分层 k 均值	0.112	0.101	0.032
			0.176	0.174	0.144
		k-中心点	0.036	0.016	0.001
		纳米纤维	0.116	0.100	0.033
	unwtd word2vec	分层的	0.086	0.079	0.027
		k-均值 k-中	0.144	0.142	0.114
		心点	0.020	0.013	0.008
		纳米纤维	0.089	0.071	0.015
	TF-IDF	层次结构	0.009	0.003	0.000
		k-均值	0.005	0.004	0.002
		k-中心点	0.008	0.000	0.000
		纳米纤维	0.006	0.005	0.000
	[illegible]	LDA	0.008	0.007	0.007

标签主题标签。在 #Auspol 数据中,我们的地面真相主题标签是前 13 个不同的主题标签,这些标签从
特征生成和聚类之前的文本。因此,这些标签可以被视为潜在标记。我们首先确定了顶部
每个集群按频率排列的三个主题标签（主题标签）。为了进行比较,我们使用所有数据从原始数据创建了一个tf-idf矩阵
主题标签,包括主题主题标签,并排除所有其他标记。然后我们提取了tf-idf最高的三个主题标签
我们对每个集群的得分进行分析,并与排名前三的主题标签进行比较。表 11概述了结果。排名靠前的主题与
每个集群的热门标签。在 13 个集群的 39 个热门主题中,只有 7 个包含不同的标签（以粗体文本突出显示）。
还有两个集群的顺序进行了调整。我们的结论是 doc2vec 聚类准确地捕获了
潜在标签标签的结构。

查看聚类质量的另一种方法是分析真实标签和聚类之间的重叠。在里面
出于空间兴趣,我们考虑了仅包含 5 个主题的 Reddit 数据集,并选择了文档大小为
101 到 200 个字符之间,以与 Twitter 数据集保持一致。然后,我们分析了 doc2vec
使用 k-means 聚类的特征与真实标签（subreddit 页面)进行聚类。结果如表12所示,可见
第一个集群将来自 subreddit 页面“NFL”的大部分父评论分组在一起,第二个集群将强烈
围绕“pcmasterrace”的页面。这些页面明显代表不同的主题。集群 3 和 4 很好地围绕“新闻”和“电影”分组
但第 5 簇主要分为“关系”和“新闻”。

为了进一步解释这个 Reddit 数据集的主题,我们按簇分析了最热门的单词。对于每个簇,我们计算了
质心作为集群中每个文档的 doc2vec 表示的平均值。由于经过训练的 doc2vec 模型产生了
文档嵌入与词嵌入在同一空间中,我们计算了簇质心和
的话。其背后的想法是,靠近聚类质心的单词可能代表该聚类。然而,这
该方法没有考虑每个簇中出现的单词的频率,或者跨簇的单词的相对频率
集群。为了整合这些信息,我们汇集了每个集群中的所有文档并计算了一个tf-idf矩阵。然后我们创建了一个
根据余弦相似度和tf-idf得分之和,为每个单词和聚类计算综合得分。表 13显示了排名前 10 的单词
按此方法排序的每个簇。很明显,该方法提取了与 Reddit 主要子页面相关的非常具体的术语,
特别是对于集群 1 和集群 2。

表 10

按文档长度范围 201-500 和 501 或更大对每种方法的 Reddit 数据进行性能评估。

文档字符长度	特征表示	聚类	神经肌肉瘤	急性心肌梗死	阿里
201–500	文档2vec	分层 k 均值	0.261	0.254	0.212
			0.487	0.483	0.496
		k-中心点	0.037	0.010	0.002
		光纤纳米颗粒纤维	0.194	0.128	0.044
	wtd word2vec	分层 k 均值 k 中	0.265	0.246	0.142
		心点	0.333	0.331	0.276
			0.174	0.172	0.150
		纳米纤维	0.247	0.226	0.133
	unwtd word2vec	层次结构	0.227	0.200	0.084
		K 均值	0.303	0.301	0.247
		k-中心点	0.106	0.103	0.081
		光纤纳米颗粒纤维	0.208	0.183	0.092
	TF-IDF	分层 k 均值 k 中	0.103	0.061	0.015
		心点	0.095	0.085	0.044
			0.014	0.013	0.007
		光纤纳米颗粒纤维	0.062	0.057	0.046
	激光衍射分析	激光衍射分析	0.080	0.079	0.071
501 +	doc2vec	分层的	0.532	0.518	0.499
		K 均值	0.686	0.684	0.708
		k-中心点	0.094	0.037	0.007
		纳米纤维	0.331	0.255	0.154
	wtd word2vec	分层 k 均值	0.465	0.400	0.327
			0.461	0.433	0.403
		k-中心点	0.353	0.330	0.283
		纳米纤维	0.366	0.325	0.229
	unwtd word2vec	层次结构	0.416	0.367	0.306
		k 均值 k 中	0.433	0.405	0.385
		心点	0.336	0.322	0.290
		纳米纤维	0.290	0.242	0.159
	TF-IDF	层次结构	0.304	0.244	0.199
		公里平均值	0.431	0.382	0.323
		类风湿关节炎	0.042	0.007	0.001
		光纤纳米颗粒纤维	0.396	0.344	0.299
	LDA	激光衍射分析	0.341	0.326	0.291

表11

每个集群的前三个主题标签和前三个主题标签。请注意，主题标签未出现在聚类数据中，但主要是
当我们为按集群汇集的推文创建tf-idf矩阵并选择得分最高的三个主题标签时，按顺序恢复。
前三个主题标签和前三个主题标签之间的差异以粗体突出显示。

簇	前三个主题标签	前三个tf-idfscore主题标签
1	#nbn,#lnp,#insiders #uspoli、	#nbn,#lnp,#insiders
2	#insiders、 #turnbull #insiders,#lnp、	#uspoli,#insiders、 #trump
3	#qldpol #qldpol,#insiders,#lnp	#内部人士,#lnp,#qldpol
4	#politas、 # qldpol、 # lnp	#qldpol,#insiders,#lnp
5	#qldpol,#qanda,#trump	#politas、 #utas,#发现
6	#insiders,#lnp,#qldpol #qldpol、	#qldpol、 #politas、 #qanda
7	#stopadani,#springst #lnp、	#内部人士,#lnp,#qldpol
8	#trump,# uspoli #qanda,#insiders、	#qldpol,#stopadani,#springst
9	#qldpol #marriageequality、	#LNP,#特朗普， #内部人士
10	#politas、 #lnp #qldpol,#stopadani、	#qanda,#insiders、 #sayitwithstickers
11	#qanda #climatechange,#qldpol,#stopadani	#marriageequality、 #equalitycampaign、 #politas
12		#qldpol,#qanda,#stopadani
十三		#气候变化,#qldpol,#stopadani

表 12

使用 k-means 聚类方法对 Reddit 数据进行 doc2vec 表示的混淆矩阵，大小范围在 101 到 200 个字符之间。

Subreddit 页面	集群 1	集群2	集群 3	集群 4	集群 5
国际单机锦标赛	第1351章	60	298	273	395
电脑大师赛	78	1295	215	185	260
消息	93	89	952	204	538
电影	89	50	152	767	226
人际关系	32	37	116	48	557

表 13
基于嵌入空间中的组合嵌入相似度得分和tf-idf得分,每个簇的前 10 个单词。

簇	热门主题	前 10 个词
1	====	人才、弗拉科、四分卫、tds、sb、wrs、名册、海豚、滑车、foles
2	电脑大师赛	安装、ps4、r9、主板、gpu、i5、操作系统、msi、处理器、华硕
3	消息	联邦、过失杀人、地区、凶杀、经济、伊斯兰国、中国、劳工、上部、托克
4	电影	复仇者联盟,乔斯,恐怖,阿诺德,电影摄影,重看,澳大利亚,doof,胸部,mcx
5	人际关系	辱骂, 心态, 反应, rdj, xanax, 婚姻, 天堂, 会议, 部分, 主观

5. 讨论

通过本研究可以清楚地看出,对于将 OSN 文本数据聚类到主题中,通常使用 doc2vec 特征表示与任何其他方法相比,与 k 均值聚类相结合可提供最佳性能。然而,这种情况方法效果不佳,需要讨论。在 RepLab 2013 Twitter 数据集上,doc2vec 方法给出了性能低于 NMI 和 AMI 测量的平均 word2vec 方法,但在 100 个 epoch 之后对 ARI 给出了最佳性能训练。此外,未加权平均 word2vec 方法的表现优于tf-idf加权平均 word2vec 方法就这个数据而言。这两个结果都不同于其他两个数据集的结果。 #Auspol 和 Reddit 上的结果文档长度在 101 到 200 个字符之间的数据表明,问题不在于每个文档的大小RepLab 数据,但很可能使用的数据量不足以准确训练 doc2vec 模型。这这意味着 doc2vec 模型应该在大于 3000 左右的数据量上进行训练。有趣的是,Reddit长度在 101 到 200 个字符之间的数据仅包含 8360 个文档,doc2vec 表现非常好,尽管Reddit 评论在使用术语上可能与推文有很大不同。

另一个有趣的观察是,在 #Auspol Twitter 数据上,带有 NMF 的tf-idf矩阵在NMI 和 AMI 的测量结果优于两种 word2vec 方法的最佳聚类,尽管 ARI 的得分较低。 2013 年 RepLab 上数据上,word2vec 方法在 NMI 和 AMI 上表现更好,但tf-idf方法在 ARI 上表现非常接近。然而,在 Reddit数据中， tf-idf方法的性能非常低,直到文档大小超过 200 个字符。这表明主题来自 Twitter 的文本可能严重依赖关键字,因为tf-idf聚类表现相对较好,考虑到这一点并不奇怪使用用户提及和主题标签。 doc2vec 方法在 #Auspol 数据上比其他特征方法。为 doc2vec 模型的主题标签和用户提及分配更大的权重可能会带来改进Twitter 数据上的表现。

根据 Reddit 数据,这项研究得出了两个有用的结果。首先，doc2vec 与文档的平均长度成反比。这一结果为未来使用 doc2vec 进行研究提供了一些指导OSN 数据。不幸的是,这个结果与 #Auspol 数据的结果不一致,这可能是由于主题标签造成的本身并没有明显的区别。使用 Twitter 数据面临持续的挑战,因为手动标记主题已经是时候了耗时且容易出错,可检索的推文数量会随着时间的推移而减少。结果与 RepLab 的2013 年 Twitter 数据,但正如已经讨论过的,数据量很小。第二个结果是doc2vec的性能方法随着文档长度的增加而增加。该方法对最长的 Reddit 评论表现出色,因此应该一般而言,应用于来自 OSN 平台的文本数据时会得到良好的结果。

改进 OSN 文档的嵌入表示法对多种自然语言处理任务非常有用。此类文档级别的重新表示法可以提供高质量的特征矩阵,供其他机器学习系统使用。

示例应用程序用于情感分析（Lee、Jin 和 Kim,2016）。此外,之前已经表明,预训练doc2vec 使用的词向量可以提高多种自然语言处理任务的性能（Lau & Baldwin,2016）。针对大量 OSN 数据预训练词向量和文档向量可以提高应用程序的性能

专注于特定的数据样本。例如,预训练的文档向量可用于流式文档分类或聚类应用。此外,这种方法还可以应用于其他领域,在这些领域中,数据可以建模为具有少量代币。例如,嵌入模型正在电子健康记录数据上得到应用（Choi 等人，2016）。在这种情况下,医疗代码被视为令牌,然后可以使用嵌入模型来捕获有关的信息疾病和治疗之间的关系,并用于后续的预测或聚类任务。

6. 结论和未来工作

在本研究中,我们展示了几种文档聚类和主题建模方法在社交媒体上的不同性能文本数据。我们的结果表明,在线社交网络数据的文档和词嵌入表示可能有效地用作文档聚类的基础。这些方法优于传统的基于tf-idf的方法和主题建模技术。此外,doc2vec 和tf-idf加权平均词嵌入表示法比文档聚类任务中词向量的简单平均值。我们还证明了 k-means 聚类提供了doc2vec 嵌入的最佳性能。

通过将这些方法应用于按文档长度范围划分的 Reddit 数据集,我们概述了以下两个关键结果：聚类 doc2vec 嵌入。首先,最佳训练周期数通常与字符成反比文档的长度范围。其次,使用 k-means 聚类的 doc2vec 嵌入在所有方面都提供了良好的性能

文档长度范围在所使用的 Reddit 数据中。这些结果表明该方法在大多数 OSN 文本数据上应该表现良好。

为了解释这些方法得出的聚类结果,我们开发了一种基于tf-idf分数和词向量相似度相结合的热门词分析。我们证明了该方法可以为主题聚类提供一组具有代表性的关键词。我们还证明了 doc2vec 嵌入与 k-means 聚类可以成功恢复 Twitter 数据中的潜在主题标签结构。

我们计划对这项工作进行多次扩展。首先,doc2vec 嵌入与 k 均值聚类相结合,可以轻松应用于任何社交媒体文本数据。在进一步的应用中,我们打算展示该方法在以流方式定义和解释动态主题方面的有用性。其次,该方法可以扩展以合并社交网络中可用的附加数据,特别是来自 Twitter 用户和网络数据。第三,神经嵌入和深度学习技术应用的最新进展,例如上下文嵌入模型(Peters et al., 2018)、潜在 LSTM 分配(Zaheer, Ahmed, & Smola, 2017)和基于深度学习的聚类模型(Min et al., 2018)可用于提供改进的特征表示或文档聚类。单词和文档嵌入也可以用作深度聚类和主题建模技术中的预训练初始层。

致谢和声明

这项研究没有获得公共、商业或非营利部门资助机构的任何具体资助。

补充材料

与本文相关的补充材料可以在在线版本中找到,网址为doi:10.1016/j.ipm.2019.04.002。

参考

Alghamdi, R. 和 Alfalqi, K. (2015)。文本挖掘中主题建模的调查。国际高级计算机科学与应用杂志， 3(7), 774–777。
Alnajran, N.、Crockett, K.、McLean, D. 和 Latham, A. (2017)。Twitter数据的聚类分析:算法回顾。第 9 届代理和人工智能国际会议论文集 - 第 2 卷:ICAART,INSTICC。SciTePress239–249 。

Amigó, E.、Carrillo de Albornoz, J.、Chugur, I.、Corujo, A.、Gonzalo, J.、Martín, T. 等。 (2013)。 RepLab 2013 概述:评估在线声誉监控系统。
CLEF 倡议第四届国际会议论文集 333–352。
Bakshy, E.、Rosenn, I.、Marlow, C. 和 Adamic, L. (2012)。社交网络在信息传播中的作用。第21届世界范围国际会议论文集
网络系列 WWW 12。美国纽约州纽约:ACM519–528。
Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003)。神经概率语言模型。机器学习研究杂志， 3,1137–1155。
Billah Nagoudi, EM、Ferrero, J. 和 Schwab, D. (2017)。 SemEval-2017 上的 LIM-LIG 任务 1:通过向量加权增强阿拉伯语句子的语义相似性。
国际语义评估研讨会（SemEval-2017)第 11 届国际语义评估研讨会（SemEval-2017)论文集125–129加拿大温哥华

Bisht, S. 和 Paul, A. (2013)。文档聚类:回顾。国际计算机应用杂志,73,26-33 。
Blei, DM, Ng, AY 和 Jordan, MI (2003)。潜在狄利克雷分配。机器学习研究杂志， 3,993–1022。
Chinnov, A.、Kerschke, P.、Meske, C.、Stieglitz, S. 和 Trautmann, H. (2015)。通过社交媒体分析在 Twitter 通信中发现主题的概述。
美洲信息系统会议论文集1-10。
Choi, E.、Bahadori, MT、Searles, E.、Coffey, C.、Thompson, M.、Bost, J. 等人 (2016)。医学概念的多层表征学习。第 22 届 ACM 论文集
SIGKDD 知识发现和数据挖掘国际会议。美国纽约州纽约:ACM1495–1504。
Corrêa Júnior, EA、Marinho, VQ 和 dos Santos, LB (2017)。NILC -USP 在 SemEval-2017 任务 4 中的应用:用于 Twitter 情绪分析的多视图集成。第 11 届国际语义评估研讨会论文集 (SemEval-2017)。计
算语言学协会 611–615。
Curiskis, S.、Drake, B.、Osborn, T. 和 Kennedy, P. 已提交。来自 Twitter 和 Reddit 的带主题标签的在线社交网络数据集。简要数据。
Dhingra, B.、Zhou, Z.、Fitzpatrick, D.、Muehl, M. 和 Cohen, W. (2016)。Tweet2vec :基于字符的社交媒体分布式表示。第 54 届
计算语言学协会年会（第 2 卷:短论文）。计算语言学协会269-274。
Fang, Y., Zhang, H., Ye, Y., & Li, X. (2014). 从 Twitter 中检测热门话题:一种多视图方法。信息科学杂志， 40(5),578–593。
Ferrara, E.、JafariAsbagh, M.、Varol, O.、Qazvinian, V.、Menczer, F. 和 Flammini, A. (2013)。社交媒体中的模因聚类。 2013 年 IEEE/ACM 会议论文集
社会网络分析与挖掘进展国际会议。美国纽约州纽约:ACM548–555。
Godfrey, D.、Johns, C.、Meyer, CD、Race, S. 和 Sadek, C. (2014)。文本挖掘案例研究:从世界杯推文中解读推特数据。CoRR , 1–11 abs/
1408.5427
Godin, F.、Vandersmissen, B.、De Neve, W. 和 Van de Walle, R. (2015)。多媒体实验室 @ ACL WNUT NER 共享任务:使用 Twitter 微帖子的命名实体识别
分布式词语表示。嘈杂用户生成文本研讨会论文集。计算语言学协会 146–153。
Guille, A., Hacid, H., Favre, C., & Zighed, D. (2013)。在线社交网络中的信息传播:一项调查。ACM SIGMOD 记录,42,17-28 。
Gutman, J., & Nam, R. (2015). Reddit 帖子的文本分类技术报告。纽约大学。
Ha, T.、Beijnon, B.、Kim, S.、Lee, S. 和 Kim, JH (2017)。通过动态主题建模检查用户对智能手表的看法。远程信息处理和信息学， 34(7)，
1262–1273。
Hong, L. 和 Davison, BD (2010)。Twitter主题建模的实证研究。第一届社交媒体分析研讨会论文集。美国纽约州纽约:ACM80–88。
Irfan, R.、King, CK、Grages, D.、Ewen, S.、Khan, SU、Madani, SA 等。 （2015）。社交网络中文本挖掘的调查。知识工程评论， 30(2)，
157170。
JafariAsbagh, M.、Ferrara, E.、Varol, O.、Menczer, F. 和 Flammini, A. (2014)。社交媒体流中的模因聚类。社交网络分析与挖掘， 4(1), 237。
Klein, C., Clutton, P., & Polito, V. (2018)。主题建模揭示了在线阴谋论坛中的不同兴趣。《心理学前沿》， 9,1-12。
Klinczak, M., & Kaestner, C. (2016)。Twitter 主题识别聚类算法比较。拉丁美洲计算杂志 - LAJC， 3，
19–26。
Lau, JH 和 Baldwin, T. (2016)。对doc2vec 的实证评估,以及对文档嵌入生成的实用见解。第一届 doc2vec 研讨会论文集
NLP 的表征学习。计算语言学协会78-86。
Le, QV 和 Mikolov, T. (2014)。句子和文档的分布式表示。第 31 届机器学习国际会议论文集,ICML 2014，
中国北京,2014 年 6 月 21-26 日1188-1196。
李 S.、金 X. 和金 W. (2016)。使用 doc2vec 和 jst 对未标记数据集进行情感分类。第18届国际电子学年会论文集
商业:智能互联世界中的电子商务。美国纽约州纽约:ACM28:1-28:5。
Li, Q.、Shah, S.、Liu, X. 和 Nourbakhsh, A. (2017)。数据集:从推文和一般数据中学习到的词嵌入。第十一届国际会议论文集
网络和社交媒体,ICWSM 2017,加拿大魁北克省蒙特利尔,5 月 15-18 日， 2017.428-436。
Mikolov, T.、Chen, K.、Corrado, G. 和 Dean, J. (2013)。向量空间中单词表示的有效估计。 CoRR绝对/1301.3781。 1301.3781

闵英、郭X、刘Q、张G、崔J、龙J. (2018)。深度学习聚类综述:从网络架构的角度。 IEEE 接入， 6，
39501–39514。
Naik, MP、Prajapati, HB 和 Dabhi, VK (2015)。语义文档聚类调查。 2015年IEEE电气、计算机与通信国际会议
技术 (ICECCT)1-10。
Patki, U. 和 Khot, DP (2017)。文本挖掘中使用的文本文档聚类算法的文献综述。工程计算机与应用科学杂志，
6(10), 16–20。
保罗 ,MJ 和 Dredze ,M. (2014)。使用主题模型发现社交媒体中的健康主题。 《公共科学图书馆一》， 9(8) ,1-11。
Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018)。深度语境化词语表征。计算语言学协会北美分会 2018 年会议论文集:人类语言技术,第 1 卷（长篇论文）。计
算语言学协会2227–2237。

Reddit (2015)。r/datasets - 我拥有所有公开的 reddit 评论以供研究。17 亿条评论,压缩后大小为 250 GB。对此有兴趣吗？（访问日期:2019 年 1 月 19 日）。https: //www.reddit.com/r/datasets/
comments/3bxlg7/i_have_every_publicly_available_reddit_comment。 eh ek, R. 和 Sojka, P. (2010)。用于大型语料库主题建模的软件
框架。LREC 2010 研讨会论文集,主题为 NLP 框架的新挑战。
马耳他瓦莱塔:ELRA45–50。
Romano, S.、Vinh, NX、Bailey, J. 和 Verspoor, K. (2016)。调整机会聚类比较措施。机器学习研究杂志， 17(1)，
4635–4666。
Shabunina, E. 和 Pasi, G. (2018)。基于图的社交媒体流中表情识别和跟踪方法。基于知识的系统 ,139 ,108–118 。
Steinskog, A.、Therkelsen, J. 和 Gambäck, B. (2017)。通过推文聚合进行 Twitter 主题建模。第 21 届北欧计算语言学会议论文集。
计算语言学协会77-86。
Stieglitz, S.、Mirbabaie, M.、Ross, B. 和 Neuberger, C. (2018)。社交媒体分析在主题发现、数据收集和数据准备方面面临挑战。国际信息管理杂志 ,39 ,156–168 。

Suri, P. 和 Roy, NR (2017)。 LDA 和 NMF 在大型文本流数据事件检测方面的比较。 2017年第三届计算智能通信技术国际会议（CICT)1-5。

Vinh, NX, Epps, J., & Bailey, J. (2010). 聚类比较的信息理论测量:变体、属性、规范化和偶然校正。
机器学习研究杂志 ,11 ,2837–2854 。
Yang, X., Macdonald, C., & Ounis, I. (2017). 使用词向量进行推特选举分类。信息检索， 21(2–3) ,183–207。
Zaheer, M., Ahmed, A., & Smola, AJ (2017)。潜在 LSTM 分配:序列数据的联合聚类和非线性动态建模。第 34 届机器学习国际会议论文集70。第 34 届机器学习国际会议论文集国际会议中心 ,澳大利亚
悉尼:PMLR3967–3976 机器学习研究论文集。

Zhao, J., Lan, M., & Tian, JF (2015).使用传统相似度测量和词嵌入进行语义文本相似度估计。第九届国际研讨会
语义评估（SemEval 2015)117。