SLUJST

SuleLamidoUniversityJournalofScience&Technology第4卷第1&2期 [2022年7月],第27-33页https://doi.org/10.56471/slujst.v4i.266

在线ISSN:2736-0903
印刷ISSN:2736-089X
www.slujst.com.ng

情绪分类:文本向量化方法回顾:Bag of Words、Tf-Idf、Word2vec 和 Doc2vec

海萨尔·达乌达·阿布巴卡尔1 , 马哈茂德·奥马尔2 *, 穆罕默德·阿卜杜拉希·巴卡勒3

1 尼日利亚古梅尔吉加瓦州立教育学院计算机科学系。

2 部门尼日利亚索科托索科托州立大学理学院计算机科学系。

3 尼日利亚索科托 Wurno 农业与动物科学学院综合研究系。ab ubakar1986@graduate.utm.my1 、mahmoodumar24@gmail.com2 *、 bakalemuhammad @gmail.com3

抽象的

在情绪分析中,有三 (3) 种方法,即机器学习、基于词典的方法和基于规则的方法。本研究调查了涉及文本向量化或词嵌入的机器学习方法 - 这是自然语言处理任务中的一个重要步骤,因为大多数机器学习算法都使用数字输入。文本向量化涉及将语料库的单词或文档表示或映射到数字或实数的数值向量。文献中有几种关于文档/文本表示的方法,但本研究将重点关注三 (3) 种常用的方法,即词袋、TF-IDF、word2vec 和 doc2vec,并尝试找出背后的原因,以便尽快审查和推荐给研究人员。对本研究的回顾表明,TF-IDF 特征向量表示通常优于其他两 (2) 种向量化方法 word2vec 和 doc2vec,特别是在书评情绪分类方面。因此建议用于未来书评数据集的研究。

关键词:情感分类、文本向量化方法-Bag of Words、Tf-Idf、Word2vec 和 Doc2vec。

1. 简介

以计算机程序和机器学习算法可以理解的格式表示文本内容是文本情感识别的重要一步,正式称为矢量化 将文本转换或编码为机器学习的数值向量。在本研究中,三种流行的矢量化方法是:情感分类采用了 TF-IDF、word2vec 和 doc2vec。

评论向量化或词嵌入是将语料库中的单词或文档表示或映射到实数的数值向量。它是自然语言处理(即语言建模)的一个子集,旨在使用可用的数字表示来表示单词或文档。大多数机器学习算法越来越多地用于文本分类和情感分析研究,旨在分析数字输入,因此这个阶段至关重要。文献中有多种表示文档/文本的方法;然而,本研究只探讨了三种广泛使用的方法,即词频、逆文档频率、word2vec 和 doc2vec 方法。

术语频率-逆文档频率(缩写为tf-idf)方法被广泛使用,因为它可以揭示文档或语料库中术语的相对重要性。维数灾难、数据稀疏性以及难以捕获文档中单词之间的语义链接都是值得注意的限制。 Word2Vec 是一种神经网络语言模型,它利用神经网络的能力来学习单词的分布式表示,以解决语义相关性和维数灾难的问题。术语频率-逆文档频率(缩写为tf-idf)方法被广泛使用,因为它可以揭示文档或语料库中术语的相对重要性。它存在许多缺陷,包括维数灾难、数据稀疏性以及难以获取文档中单词之间的数据语义关系。

SLUJST
SuleLamidoUniversityJournalofScience&Technology第4卷第1&2期 [2022年7月],第27-33页https://doi.org/10.56471/slujst.v4i.266
在线ISSN:2736-0903
打印ISSN:2736-089X
www.slujst.com.ng

Word2Vec 属于一类神经网络语言模型,它通过利用神经网络的力量来学习其分布式表示来解决单词的语义相关性和维度灾难问题。 Word2Vec 是许多近期单词和文档嵌入技术的灵感来源。

Word2vec 模型是一种具有一个隐藏层的浅层神经网络,可在大型输入语料库上进行训练后将单词映射到低维 53 向量空间。较近和较远的词向量分别反映相似和不相似词的向量空间的维度由隐藏层中的神经元数量表示。因此,许多情感分析和文本分类任务可以使用这些词向量作为特征。在word2vec中,有两种类型的神经网络模型:skipgram和连续词袋。 Skip Grams 模型使用给定输入单词的单层神经网络来预测邻近的上下文单词。通过组合多个上下文术语,连续词袋模型类似于skip-gram模型。

Doc2vec,也称为段落向量,是 word2vec 的泛化,用于容纳句子、段落和文档中的单词序列。它通过使用固定长度的密集特征向量表示文本序列,同时保留单词的顺序和语义,从而避开了基于词袋的模型的大部分弱点。

2. 相关作品

文本挖掘可以看作是知识提取,或文本数据挖掘（Team,2019）,也可以看作是数据库中知识发现（KDD)的一种方法。文本挖掘是计算机从各种书面资源中自动提取新的（未知的)信息。文本挖掘不同于在线搜索者寻找以前由其他人记录和编写的内容。由此可以清楚地看出,问题在于所有与您的需求无关的知识都应该被忽略。文本挖掘的目的是寻找无人知晓且尚未记录下来的未知信息（Gupta,2021）。

文本挖掘是数据挖掘领域的一种变化,它试图在大型数据库中找到有趣的模式（Yogapreethi & S,2016）。在非结构化文本中提取有趣且重要的信息和知识的过程,也称为智能文本分析、文本数据挖掘或文本知识发现 (KDT)。文本挖掘是一个新兴的跨学科领域,重点关注信息恢复、数据挖掘、机器学习、统计学和计算机语言学。

2.1 文本向量化文本向量化或词嵌

入涉及将语料库的单词或文档表示或映射到数字或实数的数值向量（Prabhu,2019）。由于大多数机器学习算法都处理数字输入,因此这是基于机器学习的自然语言处理任务和情感分析的必要步骤。在相关文献中,还有其他表示文档/文本的方式;然而,这里介绍了词袋、TF-IDF、word2vec 和 doc2vec 嵌入方法,因为它们与当前的主题相关。

2.2 词袋模型

词袋文本向量化模型将文档视为单词的集合,无论单词的语法或顺序如何;因此得名“词袋”。它将每个文档视为一组具有固定长度（通常是语料库中唯一单词的数量)的数值向量,每个特征代表每个单词出现的频率。 （Huspi、Abubakar 和 Umar,2021）。

例如,三个句子Sent1、Sent 2和Sent 3将分别编码,如表所示

2.1 由（Abubakar,2020年)进行:

发送 1:“Shah 是一位好作家”
发送 2:“要想成为一名优秀的作家,你需要练习”
第三个消息:“我很享受这一切”

苏雷拉米多大学科学与技术杂志第4卷第1期和第2期[2022年7月],第27-33页https://doi.org/10.56471/slujst.v4i.266

在线ISSN:2736-0903
打印ISSN:2736-089X
www.slujst.com.ng

2.3 词频-逆文档频率（TFIDF）

TFIDF 是流行的词袋游戏的缩小版。单词在文档中出现的频率（词频)与语料库中文档总数除以语料库中单词 I 出现的文档总数的对数的乘积代表文档中的每个单词 I （逆文档频率）（Trstenjaka、Mikacb 和 Donkoc,2014）。这意味着将评论文本视为文档时,将单词在评论文本中出现的次数乘以语料库中评论总数除以该术语出现的评论数量的对数。公式 1.1 是 TFIDF 的数学公式。

$$(,) \times \quad \frac{+,}{+,-} \qquad 1.1$$

其中（t,d)表示文档d中单词/术语t的原始计数,Nd表示语料库中的文档总数,Nd,t表示包含术语t的文档数量。对于 TF-IDF 向量表示,表 2.1 中的示例将使用公式 (1.1) 重新创建,如表 2.2 所示。

2.4 Word2vec

Word2Vec 是称为嵌入的文本向量化技术系列的一部分,其中使用浅层神经网络根据单词的语言上下文将单词投影到较低维的数字向量空间中。它是单词的分布式向量表示,基于这样的假设:在相同上下文中具有可比含义的单词的表示方式相似。因此,具有相似含义的词向量是

在向量空间中聚集在一起。这一系列向量化方法解决了词袋及其 n-gram 版本所特有的高维性和词上下文丢失问题。

连续词袋 (CBOW) 和 word2vec 是 word2vec 的两种实现。 （Mikolov、Sutskever、Chen、Corrado 和 Dean,2013 年)和 Skip Gram（T.、Sutskever 和 K.,2013 年）。下面解释了每个实现； 2.4.1 连续词袋（CBOW)是一种word2vec架构,可以根据一个或多个上下文词来预测当前词或目标词。在此架构中,词汇表中的每个单词

都映射到唯一的 one-hot 编码向量,并作为列存储在矩阵 W 中。对于每个唯一的单词,长度等于词汇表大小的每个编码向量中只有一个位置设置为 1,而其他位置设置为 0。 W 列中的每个单词都经过排列,使其对应于其索引或位置词汇表。

为了预测特定的目标单词,使用与给定单词组匹配的列向量的组合。
图 2.1 描述了 CBOW 中发生的事情的一个很好的例子。

2.4.2 连续 Skip Grams:CBOW 模型与此 word2vec 架构截然相反。例如,给定一个单词作为输入,该模型负责预测预定上下文单词窗口的上下文单词。在此结构中,相距较近的单词会获得更大的权重。与 CBOW 一样,词汇表中的每个单词都映射到列矩阵 W 并进行独热编码。如图 2.2 所示,唯一的变化在于模型架构,因此也在于问题的表述。

表 2.1 单词的词袋表示（Abubakar,2020）

	沙阿伊斯		a	优秀作家需要练习						I	享受每一点				It
已发送 1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
已发送 2	0	0	1	1	1	2	1	1	1	1	0	0	0	0	0
已发送 3	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

SuleLamidoUniversityJournalofScience&Technology第4卷第1&2期 [2022年7月],第27-33页https://doi.org/10.56471/slujst.v4i.266

在线ISSN:2736-0903
打印ISSN:2736-089X
www.slujst.com.ng

表 2.2词频逆文档频率表示（Abubakar,2020）

	沙H	是	一个好的作家				你需要练习吗?					享受每一点			的	它
发送1	0.48	0.48	0.18	0.18	0.18		0	0	0	0	0	0	0	0	0	0	0
发送2	0 0		0.18	0.18	0.18	0.45	0.48	0.48	0.48	0.48			0	0	0	0	0	0
发送3	0 0		0	0	0	0	0	0	0	0	0	0.48	0.48	0.48	0.48	0.48	0.48	

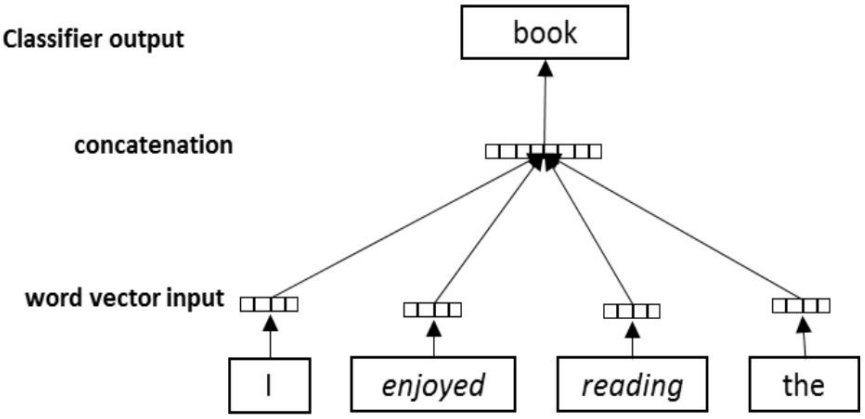

图 2.1 连续词袋模型 (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)

2.5 Doc2vec

Doc2vec 是一种基于神经的无监督学习方法,它学习不同长度的句子序列（例如段落或整个文档)的数值向量表示,如原始文件中所述
文献（Q&T,2014）。 Doc2vec 的灵感来自于 word2vec 中使用的分布式 35 个单词表示方法,其中单词向量用于通过预测随机给定的文档或段落中的下一个单词来预测句子的给定上下文中的下一个单词。我们交替使用术语“文档”和“段落”来指代句子序列。

语料库中的每一页都映射到一个唯一的 one-hot 编码向量作为列矩阵 P,此外还将词汇表中的每个单词映射到一个唯一的 one-hot 编码向量并将其存储为列矩阵 W

SuleLamidoUniversityJournalofScience&Technology第4卷第1&2期 [2022年7月],第27-33页https://doi.org/10.56471/slujst.v4i.266

在线ISSN:2736-0903
打印ISSN:2736-089X
www.slujst.com.ng

就像在 word2vec 中一样。文档向量与单词向量配对或单独使用,具体取决于 doc2vec 模型,以从段落中的预定窗口预测随机采样的固定长度上下文中的下一个单词或上下文单词。这两个 doc2vec 模型在架构上与 word2vec 类似,如下所示:

分布式内存模型（DM）（如图2.3所示)与word2vec的CBOW类似,但添加了文档向量。在反向传播神经网络训练期间,Word 与来自 P 的文档向量连接起来。虽然文档向量在所有文档之间共享,但段落向量仅在从同一段落创建的所有上下文之间共享。然后,训练后的模型可用于推断新文档的文档向量,然后将其输入到其他机器学习算法中进行预测。

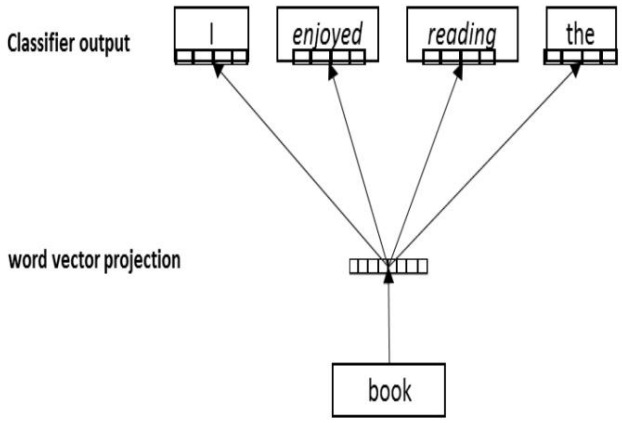

图 2.2 Skip Gram 模型（T.、Sutskever 和 K.,2013）

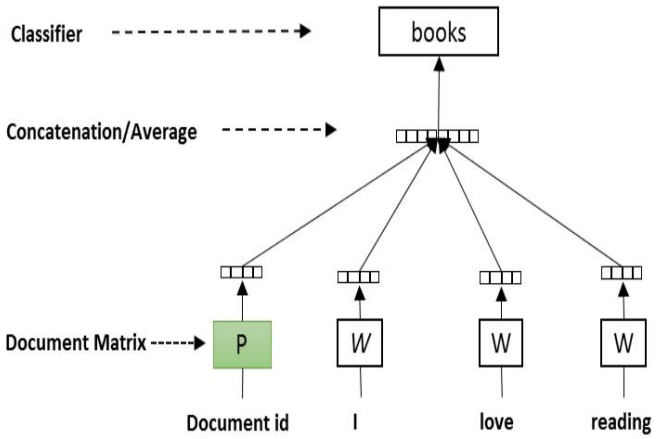

图 2.3 分布式内存模型（Q&T,2014）

3. 材料与方法

如第 2 节所述。本研究的方法涉及相关文献的收集和审查以及本研究旨在解决的问题的定义。根据以下主题搜索了 50 篇相关文献:情感分类:词向量化方法综述 - 词袋、Tf-Idf、Word2vec 和 Doc2vec。根据关键词规范缩小了搜索范围。考虑的关键词是:情感

分析、词向量化方法、词袋、tf-idf、word2vec 和 vec2word。

苏雷拉米多大学科学与技术杂志第4卷第1期和第2期[2022年7月],第27-33页https://doi.org/10.56471/slujst.v4i.266

在线ISSN:2736-0903
打印ISSN:2736-089X
www.slujst.com.ng

4.结果与讨论

之前的情感分析研究使用了 word/n-gram 包、word2vec 和 doc2vec 模型来对书面母语中的情感进行编码。这 40 种技术中的每一种都侧重于不同级别的文本粒度,这些文本粒度是通过从较低粒度级别到较高粒度级别概括一些想法来编码的。因此,信息的抽象级别由每种方法创建的特征向量表示。例如,Bag-of-words 和 n-grams 使用局部表示来关注单词和短语序列； word2vec 侧重于文本中单词和短语的代表性表示,而 doc2vec 侧重于将其上下文与整个文档相关联的单词。标准策略是单独检查这些策略中的每一种或与其他两种策略进行比较。很少有在单一情感分类研究中使用所有三种方法的情况。

5. 结论

基于对该研究主题的不同研究的回顾（情绪分类:词向量化方法回顾-词袋、Tf-Idf、Word2vec 和 Doc2vec），本研究的结果表明，TF-IDF 特征向量表示通常优于其他两 (2) 种向量化方法 word2vec 和 doc2vec,特别是在书评情绪分类方面。这可以根据 (Haisal A. D et al 2021) 对以下主题进行的研究来验证:一种使用书评数据集改进情绪分类的成对特征组合方案。与单一特征向量化方法相比,TF-IDF-word2vec、TF-IDF-doc2vec 和 doc2vecword2vec 的组合方案可以改进书评的情绪分类。这项研究还从前面提到的作者那里报告说,与所有其他方法（无论是组合还是单独)相比,TF-IDF-word2vec 的组合表现最佳。 TF-IDF 中的单词级信息与 word2vec 中的上下文信息相结合,产生了更具信息量的特征向量。此外,性能改进涵盖了四个考虑的评估指标:分类准确率、准确率、召回率和 f1-

分数。

致谢

我们要感谢马来西亚工艺大学工程学院计算机学院计算机科学系 Sharin Hazlin Bint Huspi 博士为这项工作的成功所做的贡献。

参考

阿布巴卡尔,高清 (2020)。使用书评数据集改进情感分类的成对特征组合方案。马来西亚新山:马来西亚科技大学数字图书馆。

Al-Amin, M.、Islam, MS 和 Uzzal, SD (2017)。使用 Word2vec 和单词情感信息对孟加拉语评论进行情感分析。 2017年电气、计算机和通信工程国际会议（ECCE）（第186-190页）。美国:IEEE。

Bilgin, M. 和 Senturk, IF (2017)。使用半监督 Doc2Vec 对 Twitter 数据进行情感分析。会议:2017 年国际计算机科学与工程会议 (UBMK)（第661-666 页）。UMBK:IEEE。

Chen, Q. 和 Sokolova, M.（2018 年 5 月 1 日）。 Word2Vec 和 Doc2Vec 在临床无监督情感分析中的应用。摘自https://arxiv.org/abs/1805.0035:https://arxiv.org/abs/1805.0035

Demidova, L.、Klyueva, I.、Sokolova, Y.、Stepanov, N. 和 Tyart, N. (2017)。基于SVM分类器的提高分类决策质量的智能方法。 Procedia 计算机科学， 222-230。

Doaa,ME-D。 （2016）。用于解决情感分析挑战的增强词袋模型。
国际高级计算机科学与应用杂志 (IJACSA)， 7, 1。

Gulenko, A.、Wallschlager, M.、Schmidt, F.、Kao, O. 和 Liu, F. (2016)。评估用于云中异常检测的机器学习算法。2016年 IEEE 国际大数据会议论文集， （第 2716-2721 页）。

古普塔,G.（2021 年,9 月 27 日）。 gkgupta-11 的数据挖掘简介和案例研究。取自神圣
VASTU:https://divinevastu.net/introduction-to-data-mining-with-case-studies-by-gkgupta-11/

Huspi, DS、Abubakar, HD 和 Umar, M. (2021)。使用书评数据集改进情感分类的成对特征组合方案。国际创新计算杂志， 12(1), 1。检索自。取自 https://ijic.ut

江丽、王胜、李成、 （2016）。结构扩展多项式朴素贝叶斯。信息科学。

Mikolov, T.、Sutskever, I.、Chen, K.、Corrado, GS 和 Dean, J. (2013)。单词和短语的分布式表示及其组合性。在论文集（第 26 页）。Curran Associates, Inc.

Nawangsaria, RP、Kusumaningruma, R. 和 Wibowoa, A. (2019)。Word2Vec 用于印度尼西亚语酒店情感分析。2019年第四届计算机科学与计算智能国际会议 (ICCSCI) （第360-366 页）。印度尼西亚:爱思唯尔。

田纳西州普拉布（2019 年,11 月 11 日）。了解-nlp-词嵌入-文本向量化-1a23744f7223。摘自走向数据科学:https://towardsdatascience.com/understanding-nlp-word-embeddings-text-vectorization-1a23744f7223

Q,L.,&T,M.（2014）。句子和文档的分布式表示。第 31 届国际机器学习会议 (ICML-14) 会议记录（第 1188-1196 页）。北京:科学研究。

Srujan, KS、Nikhil, SS、Rao, HR、Karthik, K. 和 Harish, BS (2018)。亚马逊书评分类基于情感分析。Nature Singapore Pte Ltd:Springer。

T., i., Sutskever, & K., IC (2013)。单词和短语的分布式表示及其组合性。
第 26 届神经信息处理系统国际会议论文集（第 3111-3119 页）。美国:Curran Associates Inc.，。

DF 团队（2019 年 10 月 14 日）。数据挖掘。取自 Data Flair:https://data-flair.training/blogs/text-mining/

Tripathy, A.、Agrawal, A. 和 Rath, SK (2016)。使用 n-gram 机器学习方法对情感评论进行分类。专家系统与应用程序，。

Trstenjaka, B.、Mikacb, S. 和 Donkoc, D. (2014)。基于 TF-IDF 的 KNN 文本分类框架。第 24 届 DAAAM 智能制造与自动化国际研讨会,2013年（第 356 – 1364 页）。
DAAM:爱思唯尔。

Yogapreethi, N. 和 S, M.（2016 年 8 月）。数据文本挖掘综述。国际软件杂志
计算 (IJSC), 7, 2/3。

Zhang, X. 和 LeCun, Y. (2015)。从 Scratch 开始理解文本。从 Scratch 开始理解文本， 1-9。