scientific reports

Check for updates

OPEN

使用多语言 BERT 对urdu文本进行多类别情感分析

Lal Khan、Ammar Amjad、Noman Ashraf 和张显宗

情感分析（SA）是一项重要任务，因为它在分析人们的观点方面发挥着至关重要的作用。然而，现有的研究仅基于英语，对资源匮乏的语言的研究有限。这项研究引入了一个基于用户评论的新的多类乌尔都语数据集，用于情感分析。该数据集收集自食品和饮料、电影和戏剧、软件和应用程序、政治和体育等各个领域。我们提出的数据集包含 9312 条评论，由人类专家手动注释为三类：正面、负面和中立。本研究的主要目标是为乌尔都语情感分析创建手动注释的数据集，并使用基于规则的机器学习（SVM、NB、Adabbost、MLP、LR 和 RF）和深度学习（CNN-1D）设置基线结果、LSTM、Bi-LSTM、GRU 和 Bi-GRU）技术。此外，我们还针对乌尔都语情感分析微调了多语言 BERT（mBERT）。我们使用四种文本表示形式：单词 n-gram、字符 n-gram、预训练的 fastText 和 BERT 单词嵌入来训练我们的分类器。我们在两个不同的数据集上训练这些模型以进行评估。研究结果表明，所提出的带有 BERT 预训练词嵌入的 mBERT 模型的性能优于深度学习、机器学习和基于规则的分类器，并取得了 81.49% 的 F1 分数。

博客、论坛、Facebook、YouTube、Twitter、Instagram 等社交网络 (SN) 最近已成为不同人群之间社交交流的最重要平台。随着技术和意识的发展，越来越多的人使用互联网进行全球交流、在线购物、分享经验和想法、远程教育以及生活各个方面的通信。用户越来越多地使用 SN 来交流他们的观点、意见和想法，以及参与讨论组。万维网 (WWW) 的不显眼性允许单个用户参与积极的 SN 语音数据，这使得文本对话，或者更准确地说，情感分析 (SA) 对于理解人们的行为至关重要。

情绪分析的重要性可以体现在我们渴望知道他们的想法以及其他人对问题的感受。公司和政府正在这些用户评论中寻找有用的信息，例如客户评论背后的感受。 SA是指应用机器、深度学习和计算语言学来调查用户撰写的评论中表达的感受或观点。由于对 SA 的兴趣日益浓厚，企业对推动营销活动、拥有更多客户、克服自身弱点以及赢得营销策略感兴趣。商业公司有兴趣了解个人对其产品和服务的反馈和看法。此外，政治家及其政党有兴趣了解他们的公众声誉。由于最近社交网络的激增，情感分析的重点已经转移到社交媒体数据研究上。 SA 的重要性在电影、戏剧、体育、新闻脱口秀、政治、骚扰、服务和医疗等多个领域中日益重要。 SA 包括 NLP 的增强技术、用于预测研究的数据挖掘以及主题建模成为令人兴奋的研究领域。

在语言学和技术方面，英语和特别是其他欧洲方言被认为是丰富的方言。然而，许多其他语言被归类为资源匮乏的语言，乌尔都语就是其中之一。乌尔都语需要一个标准的数据集，但不幸的是，学者们面临着语言资源的短缺。乌尔都语是巴基斯坦的国语，也是印度一些邦和联邦地区使用的官方语言之一。

情感分析对于乌尔都语方言和任何其他方言一样重要。许多障碍使乌尔都语的 SA 变得困难，例如乌尔都语包含正式和非正式动词形式以及阳性

1 长庚大学计算机科学与信息工程系，台湾桃园。 CIC，国立理工学院，墨西哥墨西哥城。台湾桃园长庚医院物理医学与康复科。长庚大学人工智能研究中心，台湾桃园。台湾桃园长庚大学人工智能学士课程。电子邮件：smallpig@widelab.org

以及每个名词的阴性。同样，波斯语、阿拉伯语和梵语也有乌尔都语术语。乌尔都语是从右向左书写的，单词之间的区别并不总是很清楚。由于词法问题，公认的词汇资源稀缺，乌尔都语文本数据缺乏。大多数乌尔都语网站不是采用传统的文本编码方案，而是以图解的方式组织，这使得生成最先进的机器可读语料库的任务变得复杂。众所周知的情感词典数据库是构建任何方言情感分析分类应用程序的重要组成部分。 SentiWordNet 是可用的几种英语情感词典之一。另一方面，乌尔都语是一种资源匮乏的语言，严重缺乏情感词汇。乌尔都语分词、形态结构和词汇差异等问题是开发完全有效的乌尔都语情感分析模型的主要障碍之一。

研究目的。

本研究旨在对乌尔都语评论的语义取向进行分类。我们的目标模型的灵感来自于。在引用的论文中，使用预先训练的词嵌入对阿拉伯文本进行情感分析。最近，预训练算法在 NLP 相关任务上显示出了最先进的结果。这些预先训练的模型在大型语料库上进行训练，以捕获长期语义依赖性。

本研究的目的是回答以下问题：

- 是否可以利用深度学习模型结合预先训练的词嵌入策略来识别社交网络用户用乌尔都语表达的情绪?
- 采用 fastText 和 BERT 词嵌入的深度学习方法是否比迄今为止研究的基于机器学习的方法和基于规则的乌尔都语情感分析方法更有效?

为了回答第一个研究问题，研究了使用预先训练的词嵌入对乌尔都语语言评论进行情感分析。与基于规则和基于机器学习的方法不同，基于预先训练的词嵌入的深度学习模型可以捕获单词之间的长期语义关系。为了回答第二个问题，将深度学习模型与基于机器学习的方法和基于规则的乌尔都语情感分析方法进行了比较。

我们研究的主要贡献如下：

- 基于用户评论的乌尔都语新多类情感分析数据集。它汇集了食品和饮料、电影和戏剧、软件和应用程序、政治和体育等各个领域。据我们所知，不存在这样的公共乌尔都语语料库。该语料库将公开。
- 微调用于乌尔都语情感分类的多语言 BERT 模型，该模型已针对包括乌尔都语在内的 104 种语言进行了训练，并且基于具有 12 层、768 个隐藏头和 110M 参数的 BERT 基础。
- 基于规则的方法、机器学习模型（LR、MLP、Ada-Boost、RF、SVM）和深度学习模型（1D-CNN、LSTM、Bi-LSTM、GRU 和 Bi-GRU）的一组基线结果使用不同的文本表示创建多类情感分析基准：fastText 预训练词嵌入、字符 n-gram 和单词 n-gram 特征。

本文的其余部分组织如下。 “相关工作”部分解释了情感分析的相关工作。 “语料库生成”部分描述了数据集的创建及其统计数据。 “建议的方法”部分介绍了建议的方法。 “结果分析”部分分析了实验结果和评价措施。 “结论和影响”部分对本文进行了总结。

相关工作

在本节中，我们将快速概述现有的数据集和流行的情感分析技术。

情感分析数据集。

SemEval 挑战是现有文献中为创建 SA 标准数据集而采取的最突出的努力。在每次比赛中，学者们完成不同的任务，使用不同的语料库来检查语义分析分类。此类竞赛的结果是一组标准数据集和多样化的 SA 方法。这些基准语料库是用英语和阿拉伯语创建的。用户推文/评论主要属于酒店、餐馆和笔记本电脑等各种类型。

SemEval 竞赛系列每次都会推出不同大小的语料库。 2013年，SemEval竞赛使用了SMS和Twitter语料库，Twitter语料库总共包含15,195条评论，分为训练、开发和测试数据分别为9728、1654和3813，而SMS语料库包括2093 条评论仅用于测试目的。 Twitter 语料库在 2014 年版本中总共包含 1853 条评论，其中包括 86 条用于测试的讽刺推文。 2016 年和 2017 年系列竞赛中有五个独立的子任务。每个任务的语料库分为三个部分：训练、开发和测试。子任务 A、B 和 D 以及子任务 C 和 E 分别使用了 30,632、17,639 和 30,632 个句子。 SA 的韩语语料库中有 332 篇新闻文章。人类专家手动注释这些新闻文章以进行情感分析。该数据集采用韩语主观标记语言标注方法，包含7713个主观标注句子和17615个意见表达标签，反映了韩语语言的特点。

另一个印尼语语料库已经创建。 Twitter 流 API 用于收集 350 万条推文。罗马乌尔都语语料库已创建，包含 10,021 条用户评论，属于

涉及政治、体育、食品和食谱、软件和电影等各个领域。所有这些句子都是由三位母语人士手动注释的。

情感分析方法。

现有文献中已经提出了几种方法来解决 SA 任务，例如监督和无监督机器学习。在 SemEval 2014 竞赛中，同时应用了支持向量机（SVM）和基于规则的机器学习方法。这些词典用于使用基于规则的技术来查找评论的情感极性。评论的整体极性是通过将评论中所有单词的极性分数相加并除以它们与方面术语的距离来计算的。如果一个句子的极性得分小于零 (0)，则将其分类为否定；如果分数等于0，则定义为中性；如果分数等于或大于1，则定义为正。这些分类特征和 n-gram 特征已用于训练机器学习算法。 SemEval 2016竞赛版中使用了线性回归（LR）、随机森林（RF）、高斯回归（GR）等多种机器学习算法。词嵌入是增强的自然语言处理（NLP）方法，将单词或短语表示为数字名称作为向量。 SVM 等机器学习算法将确定一个超平面，根据推文/评论的情绪对推文/评论进行分类。类似地，RF 生成各种决策树，并在做出最终选择之前检查每棵树。同样，朴素贝叶斯（NB）是一种基于贝叶斯定理的概率机器学习方法。

已经发表了许多研究报告来执行各种资源匮乏的方言，如高棉语、泰语、罗马乌尔都语、阿拉伯语和印地语。基于否定和话语关系，对印地语方言进行了情感分析。创建了印地语人工注释评论语料库。使用基于极性的方法获得了 80.21% 的准确率。同样，对泰语方言进行的研究也很少，泰方言也被认为是资源匮乏的语言。另一项研究是为了识别泰语方言中的辱骂性词语。百分之八十六的 f 测量是通过机器学习方法获得的。同样，对孟加拉语方言进行了一项研究。在本研究中，孟加拉语评论的 SA 是使用 word2vec 嵌入模型执行的。结果表明，他们提出的算法的准确率达到了 75.5%。

乌尔都语数据集和机器学习技术。

任何情感分析解决方案的基本组成部分都是计算机可读的消费者评论基准语料库。 Urdu SA 最重要的障碍之一是缺乏资源，例如缺乏乌尔都语评论的黄金标准数据集。事实是，大多数乌尔都语网站都是以说明性模式设计的，而不是使用标准乌尔都语编码。我们从现有文献中认识到两种创建数据集的方法，称为（1）自动和（2）手动。

一项专注于乌尔都语情感分析的研究创建了两个用户评论数据集来检查所提出模型的效率。 C1 数据集中仅包含 650 条电影评论，每条评论平均长度为 264 个单词。语料库 C1 中有 322 条正面评论和 328 条负面评论。另一个名为 C2 的数据集包含 700 条有关冰箱、空调和电视的评论。每条评论的平均字数为 196 个字。

另一项研究使用从 BBC 乌尔都语新闻网站收集的语料库来进行乌尔都语文本分类。成功实施了两种类型的过滤器来收集所需的数据。他们专注于“Ghusa”（愤怒）和“Pyar”（爱）等词。使用HTML解析器解析获得的数据，产生包含上述关键词的500条新闻报道和700个句子。这些句子都带有情感注释。这500篇新闻文章中，有近6000个没有情感注释的句子被丢弃。

另一项关于乌尔都语情感分析主观性的研究开发了一个语料库，其中包含来自 14 个不同领域的 151 个乌尔都语博客的 6025 个句子。三位人类专家手动将这些评论分为三类：中立、负面和正面。此外，他们还实现了五种监督机器学习算法，例如 SVM、Lib、NB（KNN、IBK）、PART 和决策树。结果表明，KNN 的准确率最高为 67.01%，并且比其他监督机器学习算法表现更好。然而，可以通过增加语料库大小并使用深度学习方法和预先训练的词嵌入模型来增强模型的性能。

同样，在工作中，NB 与 SVM 在乌尔都语文档的语言预处理步骤中的比较表明，SVM 在准确性方面比 NB 表现更好。此外，归一化的术语频率大大改善了特征选择的结果。所提出的系统的主要缺点是标记化是基于标点符号和空格完成的。然而，由于乌尔都语的语法结构，作者可能会在单个单词之间放置空格，例如（Khoubsorat，美丽），这将导致分词器将单个单词标记为两个单词（khoub）和（sorat），这是不正确的。

根据这项研究，作者使用了三种经典的机器学习算法，例如 NB、SVM 和决策树，然后采用监督机器学习方法来创建乌尔都语文本中的词义消歧 (WSD)。他们使用乌尔都语新闻网站生成的语料库来测试他们的理论。他们的 f 测量值为 0.71%。然而，通过植入自适应机制，可以提高系统的准确性。

乌尔都语数据集和深度学习技术。

最近研究了深度学习方法来对乌尔都语文本进行分类。在这项研究中，作者使用深度学习方法对产品制造的乌尔都语文档进行分类。停用词和不常用词被删除，这提高了中小型数据集的性能，但降低了大型语料库的性能。根据他们的发现，具有多个滤波器 (3、4、5) 的 CNN 优于竞争对手，而 BiLSTM 则优于 CLSTM 和 LSTM。作者使用带有多个过滤器的单层 CNN 在文档级别对文档进行分类，结果优于基线方法。对于文档分类，比较性能

语料库	公开可用	课程	算法	百分比 (%)
6025（各种流派）	Yes	3	SVM、Lib、NB、（KNN、IBK）、PART 和决策树	67
650（电影）	No	2	语言吸引人	40
700（电子电器）	No	2	语言吸引人	38
26,057 份文件	No	–	用于语言预习的 NB 和 SVM	–
乌尔都语新闻数据只有1000条意见	No	3	无监督（基于词典）	86
9601（各种域）	Yes	2	机器和深度学习	81
6000	No	2	深度学习	77.9
9312 条不同领域的评论（拟议的研究）	Yes	3	基于规则的深度学习和机器学习	78

表 1. 现有乌尔都语数据集的摘要。

领域	网站
电器、软件和博客	mobilesmspk.net, itforumpk.com, baazauq.blogspot.com, dufferistan.com, mbilalm.com, urduweb.org, urdudaan.blogspot.com, itdunya.com, achidosti. com, itdarasgah.com, tafrehmella.com,
电影、新闻脱口秀以及巴基斯坦和印度电视剧	Hamriweb.com、youtube.com、facebook.com、hamariweb.net、zemtv.com、dramasonline.com、fashionuniverse.net、tweettunnel.com
体育和娱乐	twitter.com、youtube.com、facebook.com
政治	Facebook.com、siasat.pk、twitter.com、youtube.com
食物和食谱	Urduweb.org、facebook.com、friendscorner.com、Pakistan.web.pk、kfoods。 com

表 2. 乌尔都语用户评论的在线收集来源。

混合、机器学习和深度学习模型。根据他们的发现，基于归一化差异测量的特征选择策略提高了所有模型的准确性。

在这项研究中，作者最近通过检查深度学习方法和各种词嵌入，提出了一个乌尔都语 SA 模型。对于情感分析，评估了 LSTM、BiLSTM-ATT、CNN 和 CNN-LSTM 等深度学习算法的有效性。

最近最重要的工作是使用各种机器学习和深度学习技术对乌尔都语文本进行 SA。最初，从各种社交媒体平台收集了六个不同领域的乌尔都语用户评论，以构建最先进的语料库。后来，整个乌尔都语语料库由人类专家手动注释。最后，应用RF、NB、SVM、AdaBoost、MLP、LR等一组机器学习算法和LSTM、CNN-1D等深度学习算法对生成的乌尔都语语料进行验证。 LR 算法在所有其他机器学习和深度学习算法中实现了最高的准确度。

在情感分类、概念提取和用户行为分析领域，已经开展了一些采用深度学习、语义图和多模态系统（MBS）的研究。研究中提出了一种独特的 CNN Text word2vec 模型来分析微博文本中的情感。根据测试结果，建议的 MBS 具有学习用户日常活动的正常模式并检测异常行为的卓越能力。

对乌尔都语 SA 的研究还很少，与英语等其他资源丰富的语言相比，它仍处于成熟的早期阶段。由于语言资源的稀缺，这可能会让语言工程学者感到沮丧。之前的大多数研究论文都集中在语言处理的各个领域，例如词干提取、停用词识别和删除以及乌尔都语分词和规范化。现有文献综述如表 1 所示。

此外，可用的注释数据集的大小不足以进行成功的情感分析。然而，来自有限领域的大多数数据集和评论仅来自负面和正面类别。为了解决这个问题，这项工作的重点是创建一个乌尔都语文本语料库，其中包括来自多种流派的句子。为了完成情感分析任务，我们在我们创建的语料库 UCSA-21 上应用了具有各种特征的各种机器学习模型、结合预训练词向量的深度学习模型和基于规则的算法，该语料库尚未完全研究乌尔都语情感分析文本。

语料库生成

本节介绍如何创建手动注释的乌尔都语数据集以实现乌尔都语 SA。来自多个网站的用户评论和评论的收集、人工标注规则的编写、手动标注的执行、标准化，最后是数据集特征的描述，都是创建乌尔都语情感分析语料库（UCSA- 21）。

我们从提供不受限制的访问并允许用户用乌尔都语发表评论的网站收集数据，以创建用于评估乌尔都语情绪的基准数据集。表 2 总结了我们为获取用户评论而访问的所有网站。电影、巴基斯坦和印度戏剧、电视讨论节目、食物和食谱、政治家和

Positive review examples	Negative review examples	Neutral review examples
ایک بہت عمدہ سافٹ ویئر ہے (Eik Oumda Software hai , An excellent software)	آپ کا کام بہت خراب ہے (Ap ka kaam Baoht Kharab hai Your work is very bad)	جی ہاں سپیڈ پہلے سی کم ہے لیکن چل تو رہا ہ (G haan speed pehlay se kam hai laikin chaal to raha hai, Yes, its speed is slower but at least its working)
آپ کا کام لاجواب (Ap ka Kaam lajawab, Your work is awesome)	یہ تو بہت پرانا ہو چکا ہے . (Ye to bahot porana ho chuka hai. It's too old. . . !)	اچھی ہے مگر پرانی ہو گئی ہے (Achi hai magar porani ho gai hai, Its good but old)
یہ ایک بہت ہی عمدہ نسخہ ہے (Ye Aik Bohat he umdah nuskah ha. It is a very good recipe)	بکواس (Bakwas Rubbish)	جیتنا اور ہارنا کھیل کا ایک حصہ ہے (jeetna aur harna kheil ka aik hisa, Winning and losing is part of the game)

图 1. 客户评论标记为中性、正面和负面的示例。

巴基斯坦政党、体育、软件、博客、论坛和小工具都是我们收集数据的类型。在 5 到 6 个月的时间里，三名熟悉该目标的人手动收集了用户评论。最初，数据连同以下详细信息一起收集到 Excel 表格中：(1) 评论 ID； (2) 审查范围； (3)注释标签。

为了实现 Urdu SA，我们需要一个带注释的语料库，其中包含用户评论及其情绪。
最初，定义了注释规则，然后由三位乌尔都语母语人士根据这些准则对语料库进行手动注释。所有三个以乌尔都语为母语的人都清楚注释的目的，并对完整的数据集进行了注释。根据现有文献为 Urdu SA 制定了注释指南。图 1 显示了一些来自中性、负面和正面类别的评论示例。

注释规则。

- 如果指定的评论对所有特征术语表达了积极的含义，则该评论被认为是积极的。假设它包含诸如“acha”good、“Khoubsoorat”beautiful 之类的单词，而不包含诸如“Na”“Nahi”no 之类的否定词，因为这些单词会改变极性。
- 如果任何评论表达了相互中立和积极的类别，则该评论将被标记为积极。
- 如果任何评论表达了任何同意，则该评论被归类为正面评论。
- 如果用户评论在各个方面都表达了负面情绪，那么如果该评论包含“Bora”坏、“bukwas”垃圾、“zolum”残酷、“ganda”肮脏等术语，且不包含否定作为否定，则该评论将被标记为负面颠倒整个句子的极性。
- 如果用户评论包含比任何其他类别更多的负面词，则将其分类为负面评论。
- 如果一个句子包含直接的、未经软化的分歧，那么该句子被归类为否定的。
- 如果评论包含禁止、惩罚、评估和投标等词语，则该评论将被标记为负面评论。

- 如果评论包含拒绝，则该评论被标记为负面评论。
- 如果评论包含负面术语和正面形容词，则该句子将被标记为负面评论。

- 嘲笑：诸如“MashaAllah se koy to rating milli ha na hamari cricket team ko ...akhiri he sahi”（感谢上帝的恩典，我们的板球队至少获得了一些排名。可能是最后一个）这样的句子被归类为负面句子。

- 如果一个句子包含诸如“eis team ka kia banay ga”之类的问题，这支球队会发生什么？表现出沮丧被标记为负面评论。
- 如果一个事实信息出现在一个句子中，那么这个句子就被标记为中性句子?
- 如果评论中共享假设、信念或想法，则该评论被识别为中性句子。
- 如果评论中出现诸如“也许”(Shaid) 之类的词，它们将被归类为“中性”。

特征	提议的语料库	UCSA语料库
评论总数	9312	9601
正面评价	3422	4843
负面评论	2787	4758
中立评论	3103	–
最小评论长度（字数）	1	1
评论的最大长度（以字为单位）	149	–
代币总数	179,791	1,141,716
每次评论的平均令牌数	19.30	–

表 3. 拟议的和 UCSA 乌尔都语语料库的详细信息。

- 包含有关方面的负面和正面意见的评论被视为中性句子。

语料库特征。

为了创建标准语料库，三名人类专家对整个 UCSA21 数据集进行了注释。硕士毕业生对每条用户评论进行了注释；他们的母语是乌尔都语，并且非常熟悉南非。为了确保我们的注释指南正确，我们向两个注释者（X 和 Y）随机抽取了 100 条评论样本，并要求他们标记并提及哪些评论是在哪些条件下进行的。从个人角度来看，两位注释者都将这些句子分为三类之一：否定、中性和肯定。注释者 x 和注释者 y 之间的冲突评论由第三注释者 z 解决，同时牢记上述注释指南。对于整个数据集，我们使用 Cohens Kappa 方法实现了 71.45% 的注释者间一致性 (IAA)。 IAA 评分和中等评分的结果表明，手动注释规则起草充分、易于理解，并在注释阶段得到注释专家的遵循。对数据进行评估后发现，大多数分歧发生在负面和中性分类（11.60%）和正面和中性分类（12.01%）之间。表 3 和表 4 中呈现的语料库摘要显示，UCSA-21 语料库包含 9312 条乌尔都语评论，其中正面评级为 3,422 条，负面评论为 2787 条，中立评论为 3103 条。语料库 UCSA-21 的统计数据显示了类别平衡。学者们努力创建用于情感分析研究的数据集。尽管如此，大多数可用的带注释的数据集都太小，并且仅包含来自几个域的句子，而不是像 UCSA-21 那样的多个域。大多数现有语料库的另一个缺点是它们只包含两个类别：负类和正类。

提议的方法

本节包含应用机器学习、基于规则的深度学习算法和我们提出的两层堆叠 Bi-LSTM 模型的实验描述。这些算法已在我们提出的公开可用的 UCSA-21 语料库和 UCSA 数据集上进行了训练和测试。

实验数据集。

在本研究中，我们使用两个乌尔都语数据集 UCSA-21（我们提出的）和 UCSA 来验证我们提出的模型。拟议的 UCSA-21 数据集包含从不同社交媒体网站收集的 9,312 条乌尔都语评论，属于各种类型，例如食物和食谱、电影、戏剧、电视脱口秀、政治、软件和小工具以及体育。 UCSA-21 中的每条评论属于三个类别之一：用 0 表示中性，用 1 表示正面评论，用 2 表示负面评论。三级分类已在所提议的语料库上进行了实验。 UCSA 语料库共有 9601 条正面和负面用户评论，其中包含 4843 条正面评论和 4758 条负面评论。表 3 和表 4 总结了实验中使用的数据集的详细信息。

预处理。

预处理的主要目标是使用各种步骤为后续任务准备输入文本，例如拼写纠正、乌尔都语文本清理、标记化、乌尔都语分词、乌尔都语文本规范化和停用词删除。标记化是将每个一元语法从句子中分离出来的过程。文本根据标点符号和空格进行标记。停用词是任何方言的重要词，在情感分类的上下文中没有任何意义。它们都被从语料库中删除以最小化语料库大小。分段是寻找乌尔都语单词之间边界的方法。由于乌尔都语语言的形态结构，单词之间的空格不指定单词边界。因此，确定乌尔都语中的单词边界至关重要。空格省略和空格插入是与乌尔都语分词相关的两个主要问题。两个单词之间的空格省略示例，例如“Alamgeir”，通用且类似的空格插入在单个单词中，例如“Khoub Sorat”，美丽。在乌尔都语方言中，许多单词包含多个字符串，例如“Khosh bash”，这意味着幸福是具有两个字符串的一元语法。如果在打字过程中，两个字符串之间的空格以某种方式被省略，那么它将变成“Khoshbash”，这在语法和语义上都是错误的。规范化部分可以应用适当的方法来解决阿拉伯语和乌尔都语字符的正确编码问题人物。规范化将乌尔都语方言的每个字符放入指定的 uni-code 数组 (0600-06FF) 中。

特征提取。 在文本分类等 NLP 任务中，文本通常表示为加权特征向量。本研究使用了不同的 n-gram 模型；这些是为一系列单词分配概率的模型。一元语法是具有一系列单词的模型，例如“Natural”；类似地，二元模型是两个单词的序列，例如“自然语言”，而三元模型是三个单词的序列，例如“自然语言处理”。在我们的数据集中，我们研究了 n 元语法特征，例如一元语法、二元语法、三元语法以及这些 n 元语法特征的各种组合。此外，我们还研究了各种字符语法特征以获得最佳结果。最近，预训练的词嵌入方法已经在几个 NLP 相关任务中进行了实验，其性能优于现有系统。这些词嵌入模型背后的主要思想是在大量文本数据上对其进行训练，并针对特定应用对其进行微调。维基百科和 Common Crawl (CC) 数据用于训练 fastText 词嵌入模型。维基百科是最大的免费在线数据源，以 200 多种方言编写。下载并清理数据后，对模型进行训练。 CC是一个非盈利组织，抓取网络数据并免费提供数据。 fastText 经过训练可以理解 150 多种方言，包括乌尔都语。这就是我们在我们提出的研究中选择使用 fastText 词向量模型的原因。 fastText 单词到向量模型是使用 Skipgram 和连续词袋 (CBOW) 方法的扩展进行训练的。在 Skipgram 方法中，单词表示通过字符 n 元语法进行扩展。一个向量与所有 n-gram 字符相关联，与单词相关联的向量是通过将单词中的 n-gram 字符相加获得的。 类似地，CBOW 方法将单词表示为字符 n-gram 包。

分类技术。

本节详细介绍了所提出的基于规则的机器学习集、一组深度学习算法和所提出的 mBERT 模型。 KNN、RF、NB、LR、MLP、SVM 和 AdaBoost 等机器学习算法集用于对乌尔都语评论进行分类。此外，还实现了一些深度学习算法，例如 CNN、LSTM、Bi-LSTM、GRU 和带有 fastText 嵌入的 Bi-GRU。图2解释了从数据收集到分类的抽象层框架。

基于规则的方法。 包含 4728 个负面意见词和 2607 个正面意见词的纯乌尔都语词典列表已公开。图 3 详细解释了该方法的算法。最初，每个句子都被标记化，然后通过将每个标记与乌尔都语词典中可用的意见词进行比较，将每个标记分为三个类别之一。可访问的乌尔都语词典和单词用于确定用户评论的整体情绪。如果文本包含更多正面标记，则评论将被归类为正面，极性得分为 1。如果评论包含的负面标记（单词）多于正面标记（单词），则评论将被描述为负面，极性得分为 2。最后，如果评论包含相同数量的负面和正面单词，则该评论被定义为中性，极性得分为 0。

深度学习模型。 使用 keras 神经网络库 4 实现 CNN-1D、LSTM、GRU、BI-GRU、Bi-LSTM 和带有词嵌入模型 (fastText) 的 mBERT 模型等深度学习方法进行乌尔都语情感分析，以验证我们提出的语料库。本节介绍深度学习算法的技术和实验信息。 CNN-1D 主要用于计算机视觉，但它也擅长解决自然语言处理领域的分类问题。如果您打算从整个数据集的简短固定长度块中获取新属性并且特征的位置无关，则 CNN-1D 特别有用。

研究引入 GRU 来克服循环神经网络的缺点，例如使用更新和重置门机制解决梯度消失问题。更新门和重置门本质上都是向量，控制应将哪些信息传输到输出单元。 GRU 最令人兴奋的方面是，它可以经过适当的训练来长时间保留信息，而不会丢失时间戳。具有两个 GRU 的序列处理模型称为 Bi-GRU。一种是向前获取信息，另一种是向后获取信息。该双向循环神经网络中仅存在输入门和遗忘门。

LSTM 是一种循环神经网络设计，可显示最先进的序列数据发现。 LSTM 是一种捕获文本数据之间的长期依赖关系的技术。 LSTM 模型在每个时间步获取当前单词的输入，并且前一个或最后一个单词的输出创建一个输出，该输出用于馈送到下一个状态。然后使用先前状态的隐藏层（在某些情况下，所有隐藏层）进行分类。我们使用 Bi-LSTM 模型根据类别对每个评论进行分类。一般来说，Bi-LSTM 用于从之前和未来的时间序列中捕获更多上下文信息。在本研究中，我们使用两层（前向和后向）Bi-LSTM，它从 FastText 获取词嵌入。

mBERT：BERT 是当前使用最广泛的语言建模架构之一。它的泛化能力使其能够根据用户的需求修改为各种下游任务，无论是 NER 还是关系提取、问答或情感分析。图 4 显示了我们基于多语言 BERT 的提议模型的高级架构。我们使用监督训练数据微调乌尔都语情感识别的最新多语言 (mBERT) 模型。 mBERT模型基于单语言基础BERT开发，由12个transformer层和768个隐藏层组成。使用维基百科最多的前 104 种语言（包括乌尔都语）来训练 mBERT 模型。每种方言的训练数据都是从完整的维基百科转储中收集的（用户和讨论页面除外）。

变压器：BERT Small 或 Base 有 12 个变压器层，而 BERT Large 有 24 个变压器层。 Transformer 是一种自然语言处理范例，旨在执行具有远程依赖性的序列到序列活动。变压器由编码器和解码器组成。此外，编码器由两部分组成。多头注意力是第一部分，而前馈神经网络

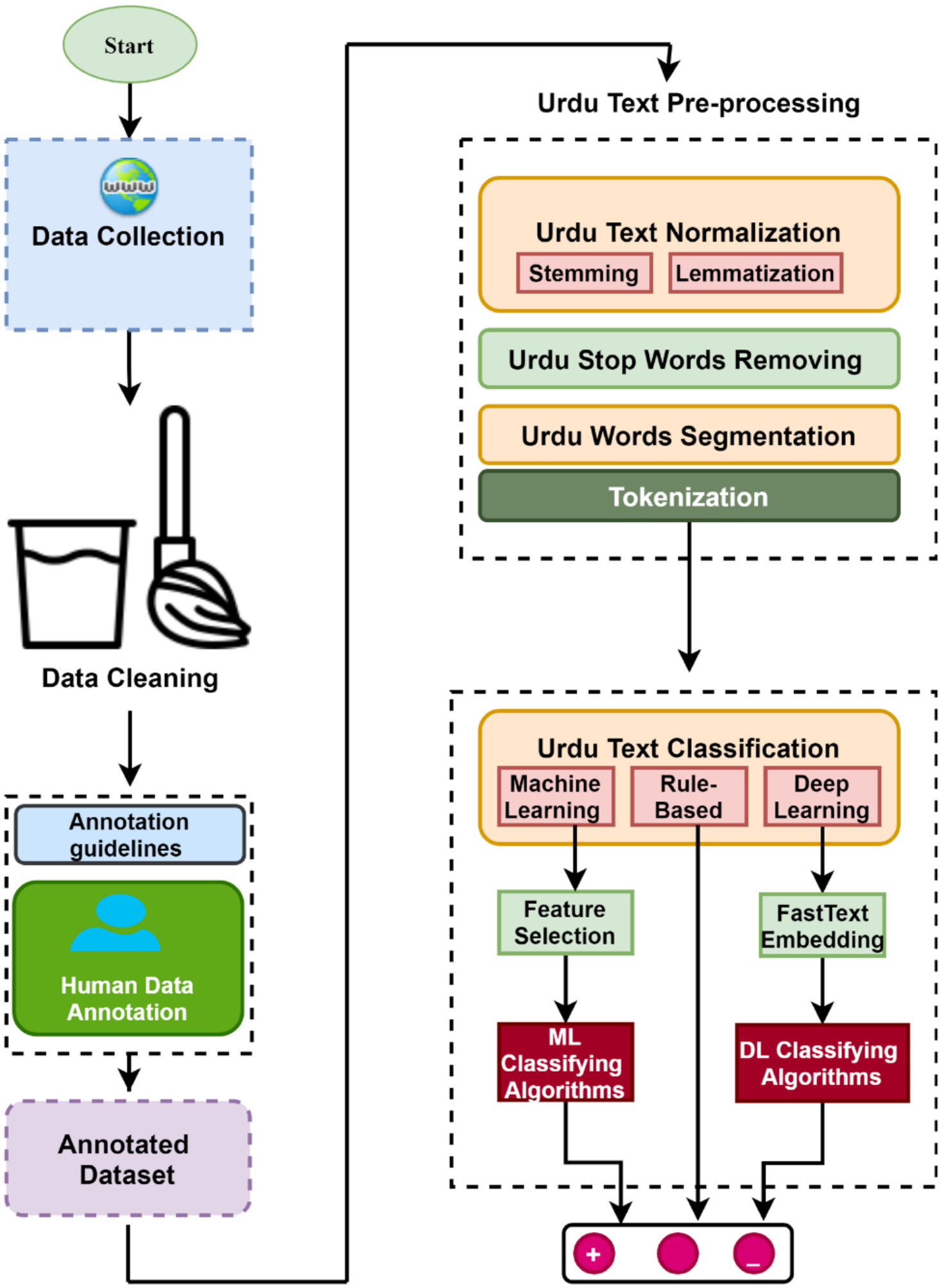

图 2. 乌尔都语情感分析的拟议抽象级别架构。

流派	总评论数	正面评价	负面评论	中立评论
食物和食谱	1250	386	317	547
电影和电视剧	1977	590	677	710
政治	1873	479	744	650
软件和小工具	2325	1326	455	544
体育及娱乐	1887	641	594	652
全部的	9312	3422	2787	3103

表 4. 建议数据集的统计数据。

Rule Based Algorithm

```
procedure  Urdu_Lexicon(args)
    Positive_Sentiment_Counter = 0
    Negative_Sentiment_Counter = 0
    Urdu_Sentiment = null
    for each word in the Urdu_Lexicon do
        if word = Positive then
            Positive_Sentiment_Counter = Positive_Sentiment_Counter + 1
        end if
        if word = Negative then
            Negative_Sentiment_Counter = Negative_Sentiment_Counter + 1
        end if
        if word is not in Urdu_Lexicon  then
            word = Neutral
        end if
    end for
    Polarity = Positive_Sentiment_Counter - Negative_Sentiment_Counter
    if Polarity < 0  then
      Urdu_Sentiment = Negative
    end if
    if Polarity > 0  then
      Urdu_Sentiment = Positive
    end if
    if Polarity = 0  then
      Urdu_Sentiment = Neutral
    end if
end procedure
```

图 3. 使用乌尔都语词典的基于规则的乌尔都语情感分析算法。

是第二部分。解码器中还包含具有多头注意前馈神经网络的屏蔽多头注意。编码器和解码器被实现为彼此堆叠。

注意力：Transformer 很大程度上依赖于注意力。 Transformers 的自注意力根据句子中的相邻单词获得文本中单词的上下文理解。注意使用等式。 (1)确定每个词的上下文。

$$\text{注意力}(Q,K,V) = \text{软最大值}\left(\frac{QK}{\sqrt{d}}\right)v \tag{1}$$

其中 Q、K 和 V 是从输入单词中提取各种组件的抽象向量。我们提出的 mBERT 模型中的特殊分类标记 将整个句子（例如“Ye tou ”）捕获到固定维度的池化表示中，并生成一个与隐藏大小和变压器输出大小相同的输出向量馈入全连接分类层，这是第一个标记的最终隐藏状态，而特殊分类标记 表示该特定句子的结尾，如图 4 所示。第二阶段是替换 15%每个句子中的标记都带有 [MASK] 标记（例如，单词“Porana”被替换为 [MASK] 标记）。然后，mBERT 模型使用非屏蔽标记的上下文来推断屏蔽标记的原始值。编码器为每个标记分配唯一的表示。例如，E1 是句子第一个词“ye”的固定呈现者。该模型由很多层组成，每个层对前一层的输出进行多头注意力，例如 mBERT 有 12 层。 T1 是中每个句子的第一个标记或单词的最后一个表示

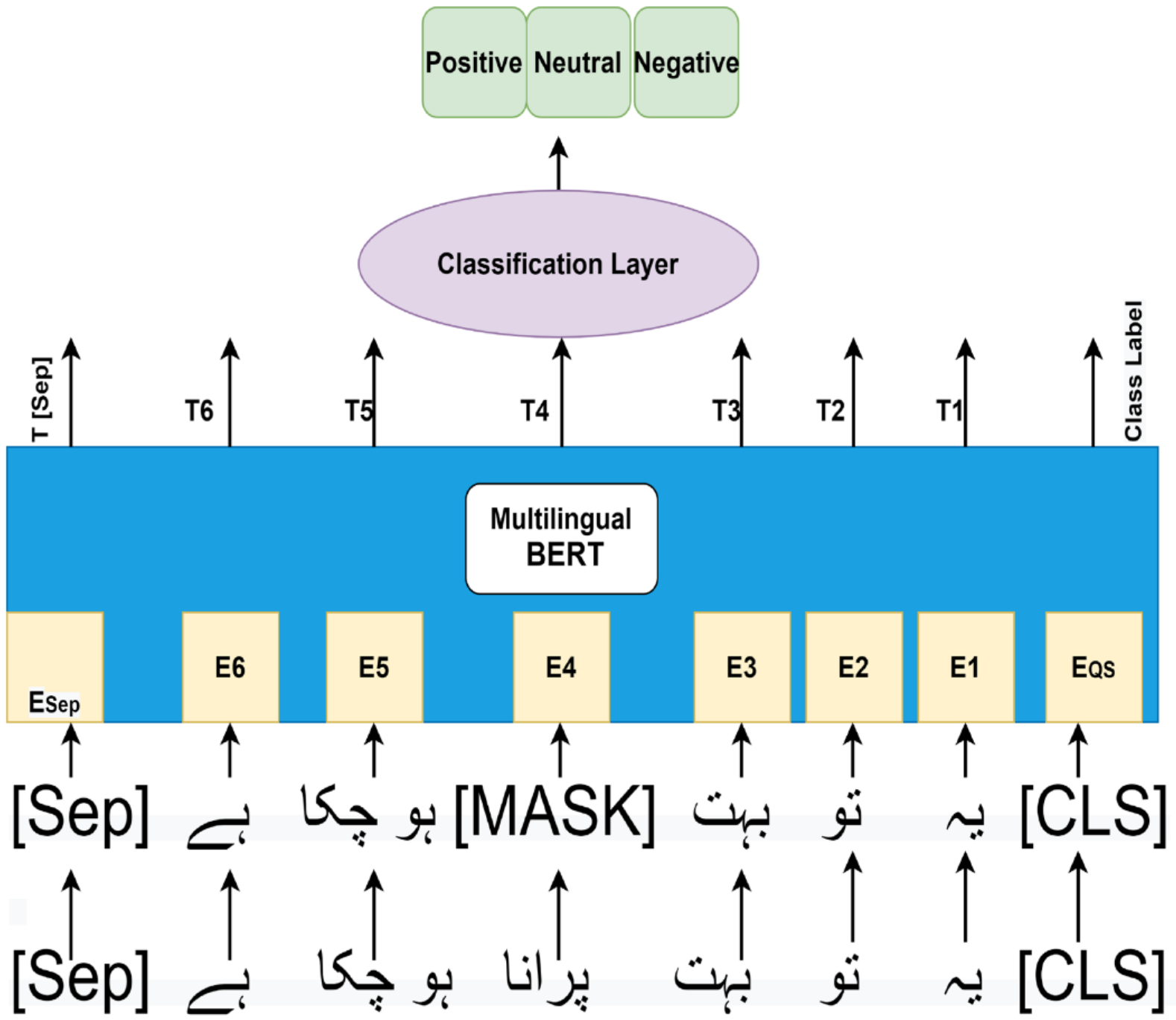

图 4. 用于乌尔都语情感分析的多语言 BERT 高级架构。

超参数	价值
学习率	2e-5
批量大小	16
纪元数	15
注意头	12
梯度累积步骤	16
隐藏尺寸	768
隐藏层	12
最大序列长度	128
参数	110M

表 5. mBERT 模型超参数。

图 4.此处添加的分类层或 softmax 层。分类层的尺寸为 K x H，其中 K 是类的数量（正类、负类和中性类），H 是隐藏状态的大小。

模型训练和微调：整个情感分类mBERT模型分两个阶段进行训练，第一阶段涉及mBERT语言模型的预训练，第二阶段涉及最外层分类层的微调。乌尔都语 mBERT 已在乌尔都语维基百科上进行了预训练。 mBERT 模型已使用提议的训练集和 UCSA 数据集进行了微调，这些数据集由带标签的用户评论组成。特别是，全连接分类层就是通过这种方式训练的。在训练过程中，使用分类交叉熵作为损失函数。表 5 列出了本研究采用的超参数。

评价措施。

在本研究中，进行了乌尔都语情感分析文本分类实验，通过使用一组机器学习、基于规则和深度学习来评估我们提出的数据集

算法。作为更好评估的基线算法，我们对建议的 UCSA-21 数据集中的 9312 条评论进行了三级分类实验。我们描述了用于评估一系列机器学习、基于规则和深度学习算法的四种评估指标，例如准确度、精确度、召回率和 F1 测量。

$$\text{准确性} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{精确} = \frac{TP}{TP + FP}$$

$$\text{记起} = \frac{TP}{TP + FN}$$

$$\text{测量值} = \frac{2 \times \text{精确} \times \text{记起}}{\text{精确} + \text{记起}}$$

其中 TN、TP、FN 和 FP 分别表示 True Negative、True Positive、False Negative 和 False Positive 的数量。

结果分析

本节解释了本研究中执行的各种实验的结果、我们为 Urdu SA 提出的架构的有用性以及对所揭示结果的讨论。在对各种已实现的机器学习、深度学习和基于规则的算法的评估中，观察到 mBERT 算法的性能优于所有其他模型。

表 6 和表 7 展示了在我们提出的 UCSA-21 语料库上使用具有不同特征的各种机器学习技术所获得的结果。结果表明，SVM 在 UCSA-21 数据集上的性能略优于其他机器学习算法，使用组合 (1-2) 特征时的准确率为 72.71%。获得的结果清楚地表明，所有机器学习分类器在使用单词特征组合（1-2）和一元组时表现更好。另一方面，获得的结果表明该组机器学习算法的性能不能满足三元词和二元词特征。使用 trigram 特征的 RF 获得 55.00% 的准确度，是所有机器学习分类器中准确度最低的。与二元词和三元词特征相比，所有机器学习分类器使用一元词特征都表现得更好，这与 一致。表 7 中列出了使用字符词组特征的几种机器学习方法的结果。使用 Char-3-gram 特征，研究结果表明，NB 和 SVM 的表现优于所有其他机器学习分类器，准确率分别为 68.29% 和 67.50%。另一方面，LR 的性能最差，在使用 char-5-gram 特征时，准确率为 58.40%。

表 8 展示了使用基于规则的方法验证我们提出的 UCSA21 数据集所获得的基线结果。基于规则的方法取得了准确率（64.20%）、精确率（60.50%）、召回率（68.09%）和F1分数（64.07）。据观察，基于规则的技术在准确率方面没有取得高分与机器学习和深度学习方法相比，基于规则的方法在该实验中表现不佳仅仅是因为实验过程中仅基于词典数据库中的术语。基于规则的算法的最大缺陷是它无法区分幽默评论和更积极的词语。诸如“MashaAllah se koy to rating milli ha na hamari cricket team ko”之类的讽刺评论被翻译为“By the ”。上帝的恩典，我们的板球队至少获得了一些排名，这可能是最后一名）”是一个负面评论，它被基于规则的方法错误地归类为正面评论。

最后，本节包含使用许多深度学习算法（例如 CNN-1D、LSTM、GRU、Bi-GRU、Bi-LSTM 和我们基于 mBERT 模型提出的模型）生成的基线结果。根据表 9 中的结果，深度学习模型优于机器学习和基于规则的方法。获得的结果表明，我们提出的基于 mBERT 和 SoftMax 进行微调的模型取代了所有其他深度学习模型，准确率、精确度、召回率和 F1 分数分别为 77.61%、76.15%、78.25% 和 77.18%。据观察，与其他传统机器学习、基于规则和深度学习算法相比，Bi-LSTM 和 Bi-GRU 可以有效地进行乌尔都语情感分析，仅仅是因为 Bi-LSTM 和 Bi-GRU 可以从后向和前向方式捕获信息。 Bi-LSTM 产生的结果稍好一些，因为它比 LSTM 和 CNN-1D 更好地理解上下文。还观察到，与最大轮询 (MP) 层相比，LSTM 和 CNN-1D 使用注意力 (ATT) 层取得了稍微更好的结果。

表 10 使用 UCSA 语料库将我们提出的 mBERT 模型的结果与其他常用深度学习算法的结果进行了比较。获得的结果表明，采用 SoftMax 的 mBERT 优于所有其他深度学习算法，准确率、精确率、召回率和 F1 分数分别为 82.50%、81.35%、81.65% 和 81.49%。我们没有应用传统的机器学习算法来验证 UCSA语料库，因为在研究中作者已经设定了基线结果。研究结果表明，由于分类类别数量较少，深度学习和我们提出的模型在使用 UCSA 语料库时表现相对更好。如上所述，UCSA 语料库仅包含两个类别：正面和负面，另一方面，我们提出的 UCSA-21 语料库包含额外的中性类别。评估数据后，在两个数据集上实现最高性能表明了我们提出的乌尔都语情感分析模型的有效性（图 5）。

混淆矩阵是评估分类有效性的度量。图 6 展示了我们使用 UCSA-21 乌尔都语语料库提出的 mBERT 的混淆矩阵。在图6中，78.10%的正面评论被正确分类为正面，而只有11.90%的正面评论被错误地分类为负面，10.00%被错误地分类为中性。在所有评论中，78.40% 的负面评论被正确识别为负面评论，而只有 11.40% 和 10.20% 的负面评论分别被错误地分类为中性和正面。仅有的

特征	模型	准确性	精确	记起	F1分数
一元语法	KNN	67.23	63.31	70.34	66.64
	RF	65.80	62.07	69.12	65.40
	NB	68.70	65.45	70.19	67.73
	LR	64.70	61.90	67.01	64.35
	MLP	67.81	65.01	70.22	67.46
	SVM	71.66	69.02	72.76	70.84
	阿达助推器	69.23	66.99	71.01	68.94
二元语法	KNN	61.73	59.21	63.04	61.06
	RF	60.58	58.97	62.10	60.49
	NB	64.39	62.05	66.20	64.05
	LR	60.24	58.10	61.98	59.97
	MLP	63.30	60.01	65.02	62.28
	SVM	67.96	64.45	69.00	66.64
	阿达助推器	64.03	61.90	66.10	63.93
卦象	KNN	58.13	48.88	68.04	57.19
	RF	55.39	47.00	67.20	55.31
	NB	59.20	51.05	70.20	59.11
	LR	55.00	47.09	65.80	54.89
	MLP	57.40	49.10	68.78	57.29
	SVM	61.66	50.00	68.10	61.25
	阿达助推器	58.50	51.01	67.80	58.21
组合 (1–2)	KNN	67.62	66.02	69.30	67.62
	RF	66.95	65.07	68.89	66.92
	NB	70.10	68.06	71.97	69.96
	LR	66.30	64.16	67.32	65.70
	MLP	69.91	67.23	70.98	69.05
	SVM	72.71	71.05	74.10	72.54
	阿达助推器	70.60	69.00	72.11	70.52
组合 (1–3)	KNN	67.80	66.80	68.33	67.55
	RF	66.70	65.70	67.32	66.50
	NB	69.50	68.44	70.12	69.26
	LR	66.00	64.70	66.39	65.53
	MLP	69.80	68.09	70.30	69.17
	SVM	71.30	70.30	72.20	71.23
	阿达助推器	71.00	69.70	71.59	70.63

表 6. 使用具有单词 n-gram 特征的机器学习模型进行乌尔都语情感分析结果。

12.00% 和 11.65% 的中性评论分别被错误分类为负面和正面，而 76.35% 的中性评论被我们提出的模型针对 UCSA-21 语料库准确分类。同样，图 7 表示我们提出的使用 UCSA 语料库的 mBERT 模型的混淆矩阵，该模型只有两类：正类和负类。

平均而言，机器学习模型包含的可训练参数少于深度神经网络，这解释了它们训练速度如此之快的原因。这些分类器不是使用语义信息，而是根据单词相对于其类别的区分能力来定义类别边界。此外，SVM 在所有采用的机器学习方法中表现得相当好，因为它不仅通过导出最大边缘超平面来处理异常值，比其他机器学习算法明显更好，而且它还支持核技术，可以有效地调整许多超平面。 -达到最佳性能的参数。此外，SVM 采用了 Hinge 损失，其性能优于 LR 的对数损失。同样，SVM 捕获特征交互的能力在一定程度上使其优于 NB，后者通常独立处理特征。

另一方面，深度学习算法不仅使特征工程过程自动化，而且比机器学习分类器更有能力提取隐藏模式。由于缺乏训练数据，机器学习方法总是不如深度学习算法成功。这正是乌尔都语情感分析实践作业的情况，其中提出的和定制的深度学习方法显着优于机器学习方法。 Bi-LSTM 和 Bi-Gru 是适应性强的深度学习方法，可以捕获向后和向前方向的信息。所提出的 mBERT 使用 BERT 词向量表示，这对于 NLP 任务非常有效。最终，这种基于变压器和编码器-解码器技术的方法击败了其他深度学习、机器学习和基于规则的模型。图 5 比较三种不同方法的总体精度

特征	模型	准确性	精确	记起	F1分数
Char-3-Gram	KNN	65.23	61.31	68.34	64.63
	RF	64.70	61.07	67.12	63.95
	NB	68.29	63.45	70.19	66.65
	LR	64.60	62.90	66.01	64.41
	MLP	66.71	63.01	68.22	65.51
	SVM	67.50	64.02	68.76	66.30
	阿达助推器	64.90	62.99	66.01	64.66
字符 4 克	KNN	60.75	59.21	62.04	60.59
	RF	60.30	57.97	60.10	59.01
	NB	63.40	60.05	64.20	62.05
	LR	60.24	57.10	60.98	58.98
	MLP	62.10	58.15	64.10	60.98
	SVM	65.90	62.35	67.10	64.63
	阿达助推器	62.90	60.70	64.20	62.40
字符 5 克	KNN	60.00	58.10	61.10	59.56
	RF	58.70	56.90	59.00	57.93
	NB	62.46	59.05	62.10	60.53
	LR	58.40	55.10	59.90	57.39
	MLP	60.10	56.01	62.00	58.85
	SVM	63.55	60.45	64.10	62.22
	阿达助推器	61.00	59.60	61.00	60.29

表 7. 使用具有 char n-gram 特征的机器学习模型进行乌尔都语情感分析结果。

模型	准确性	精确	记起	F1分数
基于规则的	64.20	60.50	68.09	64.07

表 8. 使用基于规则的算法的乌尔都语情感分析结果。

词嵌入	模型	准确性	精确	记起	F1分数
快文本	双LSTM	76.50	75.01	77.14	76.06
	双GRU	75.60	73.10	76.70	74.85
	CNN-1D	72.10	69.79	72.70	71.21
	CNN-1D+MP	70.09	68.79	70.70	69.73
	CNN-1D+ATT	73.80	71.79	75.70	73.69
	LSTM	73.15	71.40	74.28	72.49
	LSTM+MP	72.15	70.40	73.28	71.81
	LSTM+ATT	74.80	72.40	76.28	74.41
	GRU	72.50	71.00	72.00	71.49
BERT	提议的模型	77.61	76.15	78.25	77.18

表 9. 使用 UCSA-21 语料库深度学习模型的乌尔都语情感分析结果。

并提出用于乌尔都语情感分析的模型。结果表明，所提出的 mBERT 模型击败了深度学习、机器学习和基于规则的算法。

如前所述，与其他资源丰富的语言相比，乌尔都语的形态结构非常独特、极其丰富且复杂。乌尔都语是多种语言的混合体，包括印地语、阿拉伯语、土耳其语、波斯语和梵语，并包含这些语言的借词。这些是算法错误分类的最常见原因。分类不正确的其他原因包括乌尔都语文本的规范化尚不完善。要标记乌尔都语文本，必须删除/插入单词之间的空格，因为单词之间的边界不明显。同样，在乌尔都语句子中，单词的顺序可以改变，但含义保持不变，如“Meeithay aam hain”和“Aam meeithay hain”，两者都具有相同的含义“芒果很甜”。用户评论的手动标注也是漏分类的原因之一。

词嵌入	模型	准确性	精确	记起	F1分数
快文本	双LSTM	81.10	80.20	80.55	80.37
	双GRU	80.55	80.05	80.15	80.09
	CNN-1D	78.10	78.43	76.78	77.59
	CNN-1D+MP	77.60	77.05	75.25	76.13
	CNN-1D+ATT	79.05	78.00	7.45	78.15
	LSTM	78.85	77.76	77.83	77.79
	LSTM+MP	77.55	76.50	76.45	76.47
	LSTM+ATT	79.05	79.80	78.50	78.67
	GRU	78.35	77.30	77.15	77.22
BERT	提议的模型	82.50	81.35	81.65	81.49

表 10. 使用 UCSA 语料库深度学习模型的乌尔都语情感分析结果。

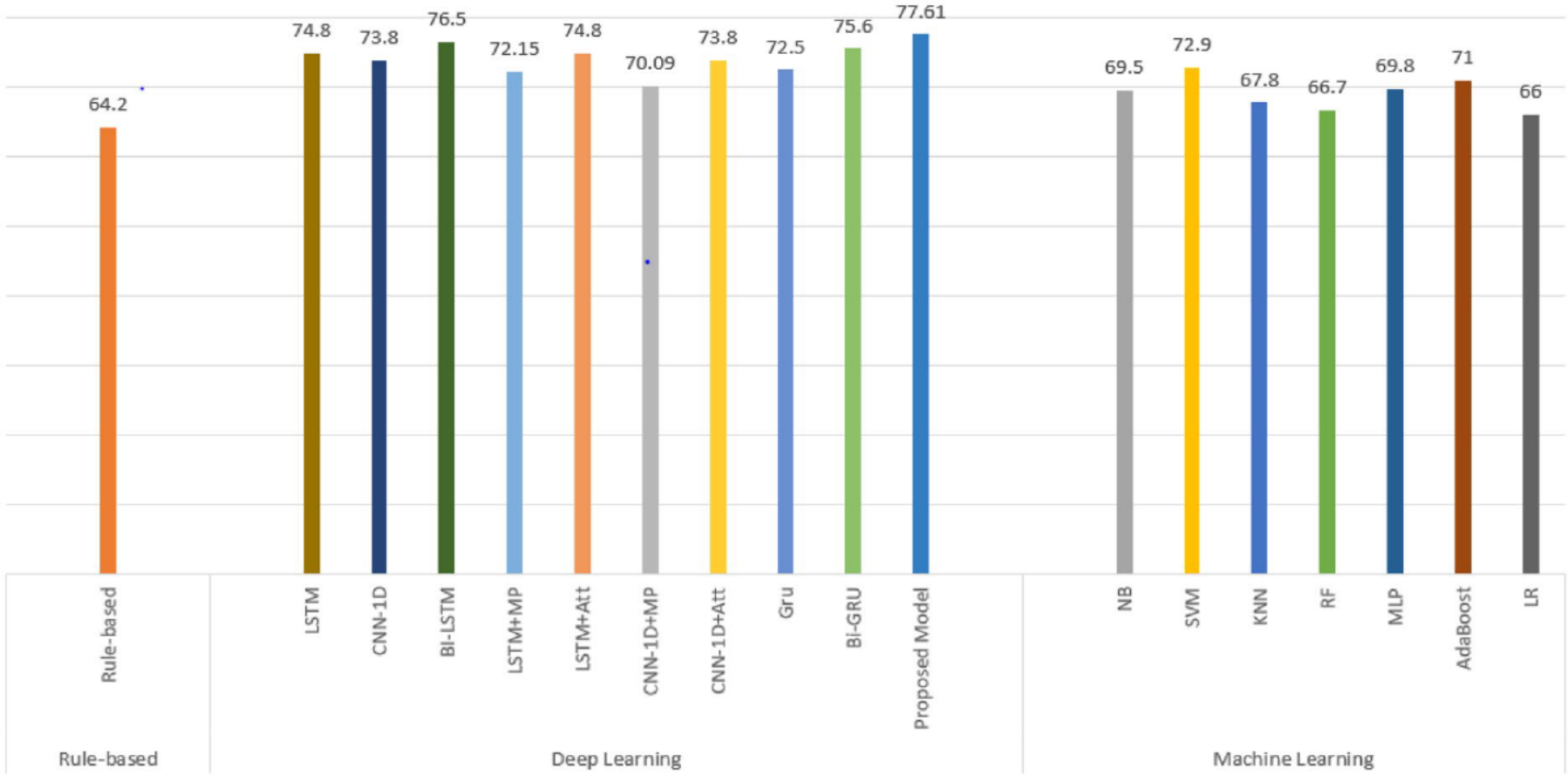

图 5. 机器、深度学习和基于规则的方法与使用 UCSA-21 语料库的建议模型的准确性比较。

Predicted/Actuals	Positive	Negative	Neutral
Positive	78.10 %	11.90 %	10.00 %
Negative	10.20 %	78.40 %	11.40 %
Neutral	11.65 %	12.00 %	76.35 %

图 6. 使用我们提出的 UCSA-21 语料库的我们提出的模型的混淆矩阵。

使用一组具有单词和字符 n-gram 特征的机器学习算法的主要目的是根据我们提出的乌尔都语语料库建立基线结果。我们提出的数据集包含短类型和长类型的用户评论，这就是为什么我们使用 GRU 和 LSTM 等各种深度学习算法来研究算法针对乌尔都语文本的性能。 GRU 通常用于对短句子进行分类，而 LSTM 由于其核心结构而被认为比长句子表现更好。同样，BERT 是目前无监督预训练性能最高的模型之一。解决屏蔽语言问题

Predicted/Actuals	Positive	Negative
Positive	83.25%	16.75 %
Negative	18.25 %	81.75 %

图 7.我们使用 UCSA 语料库提出的模型的混淆矩阵。

建模目标，该模型基于 Transformer 架构，并使用来自维基百科的大量未标记文本进行训练。它在各种 NLP 任务上表现出了出色的性能。使用 mBERT 的动机是研究其针对乌尔都语等资源匮乏语言的性能。

如前所述，使用深度学习方法来分析乌尔都语情绪的研究很少。该领域仅发表了很少的研究，并且它们都在域有限的小数据集上使用了各种机器学习分类器，并且只有正类和负类。另一方面，我们的数据集包含比早期研究更多的用户评论，并且包含具有三个分类类别的多种类型：正面、负面和中立。表 1 显示了我们的研究与以前的研究的总结和比较。

结论和启示

社交媒体平台上产生了大量数据，其中包含各种应用程序的关键信息。因此，情绪分析对于分析公众对任何产品或服务的看法至关重要。我们观察到，在乌尔都语语言中，大多数研究都集中在语言处理任务上，只有少数实验在乌尔都语情感分析领域进行，利用几种经典的机器学习方法，相对而言，数据语料库只有两个数据类。相比之下，我们提出了一个多类乌尔都语情感分析数据集，并使用各种机器和深度学习算法来创建基线结果。此外，我们提出的 mBERT 分类器使用 UCSA 和 UCSA-21 数据集分别获得了 81.49% 和 77.18% 的 F1 分数。

本文为更多深度学习研究为资源有限的语言构建独立于语言的模型奠定了道路。我们的研究结果揭示了一个重要的见解：使用预先训练的词嵌入进行深度学习是处理乌尔都语等复杂且资源匮乏的语言的可行策略。未来，我们的计划是使用 GPT、GPT2 和 GPT3 等模型来改进结果。我们相信，我们公开的数据集将作为乌尔都语情绪分析的基线。

收稿日期：2021 年 9 月 16 日；接受日期：2022 年 3 月 22 日
Published online: 31 March 2022

参考

1. 刘，Y.等人。使用在线社交网络中的异构特征识别社会角色。 J.副教授。信息。科学。技术。 70, 660–674 (2019)。
2. Lytos, A.、Lagkas, T.、Sarigiannidis, P. 和 Bontcheva, K. 论证挖掘的演变：从模型到社交媒体和新兴工具。信息。过程。管理。 56、102055（2019）。
3. Vuong, T.、Saastamoinen, M.、Jacucci, G. 和 Ruotsalo, T. 了解自然信息搜索任务中的用户行为。 J。副教授。信息。科学。技术。 70, 1248–1261 (2019)。
4. Amjad, A.、Khan, L. 和 Chang, H.-T.使用卷积神经网络的特征选择方法对语音情感分类的影响。 PeerJ 计算。科学。 7、e766（2021）。
5. Amjad, A.、Khan, L. 和 Chang, H.-T.使用具有混合特征统一的深度神经网络进行半自然和自发语音识别。流程 9, 2286 (2021)。
6. Al-Smadi, M.、Al-Ayyoub, M.、Jararweh, Y. 和 Qawasmeh, O。使用形态、句法和语义特征增强阿拉伯酒店评论的基于方面的情感分析。信息。过程。管理。 56, 308–319 (2019)。
7. Hassan, S.-U.、Safder, I.、Akram, A. 和 Kamiran, F。一种使用引文上下文分析来衡量科学知识流的新型机器学习方法。科学计量学 116, 973–996 (2018)。
8. 阿什拉夫，M.等人。本土IT行业可用性意识研究国际。 J.Adv。计算。科学。应用 9, 427–432 (2018)。
9. 沙德洛，M.等人。从科学文献中识别研究假设和新知识。 BMC 医学。通知。决定。麦。 18, 1–13 (2018)。
10. Thompson, P.、Nawaz, R.、McNaught, J. 和 Ananiadou, S. 用元知识信息丰富新闻事件。郎.资源。评估。 51, 409–438 (2017)。
11. Mateen, A.、Khalid, A.、Khan, L.、Majeed, S. 和 Akhtar, T. 控制城市车辆交通的有力算法。 2016 年 IEEE/ACIS 第 15 届计算机与信息科学国际会议 (ICIS)，1-5（IEEE，2016）。
12. Bashir, F.、Ashraf, N.、Yaqoob, A.、Rafiq, A. 和 Mustafa, R. U. 人类的攻击性和对不确定决策的反应。国际。 J.Adv。应用。科学。 6、112-116（2019）。
13. 穆斯塔法，R.U.等人。基于情感分析的社交媒体中的多类抑郁症检测。 In Latifi, S. (ed.) 第 17 届国际信息技术会议 - 新一代 (ITNG 2020), 659–662 (Springer International Publishing, Cham, 2020)。

14. Ameer, I.、Ashraf, N.、Sidorov, G. 和 Gómez Adorno, H. 使用 Twitter 中基于内容的特征进行多标签情感分类。计算。姐姐。 24, 25 (2020)。
15. 阿什拉夫，N.等人。基于 YouTube 的宗教仇恨言论和极端主义检测数据集，具有机器学习基线。 J·英特尔。模糊系统20:1-9。
16. Sailunaz, K. 和 Alhajj, R. 来自 Twitter 文本的情感和情绪分析。 J. 计算机。科学。 36、101003（2019）。
17. Khan, Z.、Iltaf, N.、Afzal, H. 和 Abbas, H。利用推荐系统的上下文嵌入丰富非负矩阵分解。神经计算 380, 246–258 (2020)。

18. Devi, B. & Pattabiraman, V. 用于在线社交网络的软余弦梯度和高斯混合联合概率推荐系统。国际。 J·英特尔。工程师。系统。 13、301311（2020）。
19. 张，B.等人。通过批评学习进行情感分析，用规则优化卷积神经网络。神经计算 356, 21–30 (2019)。

20. Luo, Z., Huang, S. & Zhu, K. Q. 知识支持从产品评论中提取突出方面。信息。过程。管理。56, 408–423 (2019)。
21. Araque, O., Zhu, G. & Iglesias, C. A. 用于情感分析的情感词典的基于语义相似性的视角。基于知识的系统165, 346–359 (2019)。

22. 萨夫德尔 (Safder)，I. 和哈桑 (Hassan)，S.-U。文献计量增强信息检索：一种新颖的深度特征工程方法，用于从全文出版物中进行算法搜索。科学计量学 119, 257–277 (2019)。
23. Al-Ayyoub, M.、Khamaiseh, A. A.、Jararweh, Y. 和 Al-Kabi, M. N. 阿拉伯语情绪分析的综合调查。信息。过程。管理。 56, 320–342 (2019)。
24. 阿斯加尔，M.Z.等人。创建用于乌尔都语情感分析的情感词典：资源匮乏语言的案例。专家系统。36，e12397（2019）。
25. Masroor, H.、Saeed, M.、Feroz, M.、Ahsan, K. 和 Islam, K. Transtech：开发罗马乌尔都语到英语的小说翻译器。Heliyon 5，e01780 (2019)。
26. Ombabi, A. H.、Ouarda, W. 和 Alimi, A. M. 使用社交网络中共享的文本信息进行阿拉伯语情感分析的深度学习 CNN-LSTM 框架。苏克。网络。肛门。分钟。 10, 1–13 (2020)。
27. Ashraf, N.、Mustafa, R.、Sidorov, G. 和 Gelbukh, A. 在线讨论中的个人与群体暴力威胁分类。 2020 年网络会议配套论文集，WWW '20, 629–633（计算机协会，美国纽约州纽约市，2020 年）。

28. Ashraf, N.、Zubiaga, A. 和 Gelbukh, A. 利用回复作为对话上下文来检测 YouTube 评论中的滥用语言。 PeerJ 计算。科学。 7、e742（2021）。
29. Amjad, M.、Ashraf, N.、Zhila, A.、Sidorov, G 和 Zubiaga, A. 乌尔都语推文中的威胁语言检测和目标识别。 IEEE 访问。 https://doi.org/10.1109/ACCESS.2021.3112500（2021）。
30. Ashraf, N.、Butt, S.、Sidorov, G. 和 Gelbukh, A. CIC 在 CheckThat！ 2021 年：使用机器学习和数据增强检测假新闻。 CLEF 2021——评估论坛会议和实验室（罗马尼亚布加勒斯特，2021 年）。
31. Kiritchenko, S.、Mohammad, S. 和 Salameh, M. Semeval-2016 任务 7：确定英语和阿拉伯语短语的情绪强度。第十届国际语义评估研讨会论文集 (SEMEVAL-2016)，42-51 (2016)。
32. Fernández, J.、Gutiérrez, Y.、Gómez, J. M. 和 Martinez-Barco, P. Gplsi：使用skipgrams 在 Twitter 中进行监督情绪分析。第八届国际语义评估研讨会论文集（SemEval 2014），294-299（2014）。
33. Jang, H.、Kim, M. 和 Shin, H. Kosac：成熟的韩国情绪分析语料库。第 27 届亚太语言、信息和计算会议记录 (PACLIC 27)，366–373 (2013)。
34. Wicaksono, A. F.、Vania, C.、Distiawan, B. 和 Adriani, M. 自动构建印尼推文情绪分析语料库。第 28 届亚太语言、信息和计算会议论文集，185–194 (2014)。
35. 马哈茂德，Z. 等人。使用循环卷积神经网络模型表达罗马乌尔都语文本中的深层情感。信息。过程。管理。57, 102233 (2020).
36. Ayata, D.、Saraclar, M. 和 Özgür, A. Busem 在 semeval-2017 任务 4a 中使用词嵌入和长期短期记忆 rnn 方法进行情感分析。第 11 届国际语义评估研讨会 (SemEval-2017) 论文集，777–783 (2017)。

37. Mittal, N.、Agarwal, B.、Chouhan, G.、Bania, N. & Pareek, P. 基于否定和话语关系的印地语评论情感分析。第 11 届亚洲语言资源研讨会论文集，45-50 (2013)。
38. Tuarob, S. & Mitrpanont, J. L. 自动发现社交网络中滥用泰语的用法。亚洲数字图书馆国际会议，267-278（Springer，2017 年）。

39. Al-Amin, M.、Islam, M. S. 和 Uzzal, S. D. 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)，186–190 (IEEE，2017)。
40. Ijaz, M. 和 Hussain, S. 基于语料库的乌尔都语词典开发。语言技术会议记录 (CLT07)，巴基斯坦白沙瓦大学，卷。 73（2007）。

41. Syed, A. Z.、Aslam, M. 和 Martinez-Enriquez, A. M. 将目标与情感单位关联起来：乌尔都语文本情感分析向前迈进了一步。阿蒂夫。英特尔。修订版 41, 535–561 (2014)。
42. Mukund, S.、Srihari, R. 和 Peterson, E. 乌尔都语（一种资源匮乏的语言）的信息提取系统。 ACM 翻译。亚洲郎.信息。过程。 9, 1–43 (2010)。
43. Mukhtar, N. 和 Khan, M. A. 使用监督机器学习方法进行乌尔都语情感分析。国际。 J.模式识别。阿蒂夫。英特尔。 32、1851001（2018）。
44. Ali, A. R. 和 Ijaz, M. 乌尔都语文本分类。第七届信息技术前沿国际会议论文集，1-7 (2009)。

45. Abid, M.、Habib, A.、Ashraf, J. 和 Shahid, A. 使用机器学习方法进行乌尔都语词义消歧。集群计算。21, 515–522 (2018)。
46. Akhter, M. P.、Jiangbin, Z.、Naqvi, I. R.、Abdelmajeed, M. 和 Fayyaz, M. 探索产品制造中乌尔都语文本分类的深度学习方法。企业信息系统。 20, 1–26 (2020)。
47. Nasim, Z. 和 Ghani, S. 使用马尔可夫链对乌尔都语推文进行情绪分析。 SN 计算。科学。 1, 1–13 (2020)。
48. 阿西姆，M.N.等人。对基于机器和深度学习的乌尔都语文本文档分类方法的性能进行基准测试。神经计算。应用。 33, 5437–5469 (2021)。
49. Naqvi, U.、Majid, A. 和 Abbas, S. A. Utsa：使用深度学习方法进行乌尔都语文本情感分析。 IEEE 访问（2021）。
50. Khan, L.、Amjad, A.、Ashraf, N.、Chang, H.-T。 & Gelbukh, A. 使用深度学习方法进行乌尔都语情感分析。 IEEE 访问（2021）。

51. 徐，D.等人。基于深度学习的微博文本情感分析信息。融合 64, 1–11 (2020)。
52. 田，Z.等人。城市大数据下的用户和实体行为分析。 ACM 翻译。数据科学。 1, 1–19 (2020)。
53. Qiu, J., Chai, Y., Tian, Z., Du, X. & Guizani, M. 基于智慧城市大数据语义图的自动概念提取。 IEEE 传输。计算。苏克。系统。 7, 225–233 (2019)。
54. Hashim, F. & Khan, M. 使用乌尔都语名词进行句子级情感分析 101–108（白沙瓦大学计算机科学系，2016 年）。

55. Do, H. H.、Prasad, P.、Maag, A. 和 Alsadoon, A. 基于方面的情感分析的深度学习：比较综述。专家系统。应用。 118, 272–299 (2019)。

56. Abdul-Mageed, M. & Diab, M. T. Awatif：用于现代标准阿拉伯语主观性和情感分析的多流派语料库。
LREC 515, 3907–3914 (2012)。
57. Maynard, D. 和 Bontcheva, K。评估社交媒体情绪分析工具的挑战。第十届国际语言资源与评估会议记录 (LREC 2016), 1142–1148
(LREC, 2016)。
58. Ganapathibhotla, M. & Liu, B. 在比较句子中挖掘观点。第 22 届国际计算语言学会议论文集（Coling 2008），241-248（2008）。

59. Mehmood, K.、Essam, D.、Shafi, K. 和 Malik, M.K. 对资源匮乏的语言罗马乌尔都语的情感分析。 ACM 翻译。
亚洲低资源。郎.信息。过程。 19, 1–15 (2019)。
60. Sorgente, A.、Vettigli, G. 和 Mele, F. 用于电影评论基于方面的情感分析的意大利语料库，349–353 (2014)。
61. Bojanowski, P.、Grave, E.、Joulin, A. 和 Mikolov, T. 用子词信息丰富词向量。跨。副教授。计算。
语言学家。 5, 135–146 (2017)。
62. Kalchbrenner, N.、Grefenstette, E. 和 Blunsom, P. 用于建模句子的卷积神经网络。 arXiv：1404.2188（arXiv 预印本）（2014
年）。
63. Rakhlin, A. 用于句子分类的卷积神经网络。 GitHub（2016）。
64. Cho，K.等人。使用 rnn 编码器-解码器学习短语表示以进行统计机器翻译。 arXiv：1406.1078（arXiv 预印本）（2014 年）。

65. Hochreiter, S. 和 Schmidhuber, J. 长短期记忆。神经计算。 9、1735-1780（1997）。
66. Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. Bert：用于语言理解的深度双向转换器的预训练。
arXiv:1810.04805（arXiv 预印本）（2018）。
67. Pires, T.、Schlinger, E. 和 Garrette, D. 多语言 bert 的多语言程度如何？ arXiv:1906.01502（arXiv 预印本）（2019）。

作者贡献

L.K.起草主要手稿文本。 H.-T.C.设定实验策略。 L.K.、A.A.、N.A. 设计并应用了这些实验。所有作者都审阅了手稿。 H.-T.C.处理流程和论文出版问题。

利益竞争

作者声明没有竞争利益。

附加信息

信函和材料请求应发送至 H.-T.C.

重印和许可信息可在 www.nature.com/reprints 上获取。

出版商说明施普林格·自然对于已出版地图和机构隶属关系中的管辖权主张保持中立。