arXiv:2304.05468v1 [cs.CL] 11 Apr 2023

塞尔维亚语自然语言处理资源与方法综述

Ulfeta Marovac[1], Aldina Avdi´c[1*]and Nikola Milo˘sevi´c[2,3]1技术与工艺科学系，诺维帕扎尔州立大学，Vuka Karad˘zi´ca 9，诺维帕扎尔，36300，塞尔维亚。2*研发，拜耳制药，M¨ullerstrasse178，柏林，13353，德国。[3] 塞尔维亚人工智能研究与发展研究所，Fru˘skogorska 1，诺维萨德，21102，塞尔维亚。

*通讯作者。电子邮箱：apljaskovic@np.ac.rs；贡献作者：umarovac@np.ac.rs；nikola.milosevic@bayer.com；

摘要

塞尔维亚语是一种斯拉夫语，有超过1,200万使用者，超过1,500万人能够很好地理解这种语言。在自然语言处理领域，它可以被视为一种低资源语言。此外，塞尔维亚语被认为是一种高度屈折语。大量的词语屈折变化与语言资源的匮乏相结合，使得塞尔维亚语的自然语言处理颇具挑战性。尽管如此，在过去三十年中，人们开展了许多开发塞尔维亚语自然语言处理资源和方法的倡议，从根据书籍和互联网中的文本开发语料库、用于分类和命名实体识别任务的标注语料库，到执行这些任务的各种方法和模型。本文回顾了这些倡议、资源、方法及其可用性。

关键词： 自然语言处理, 文本挖掘, 语言资源, 塞尔维亚语

1 引言

塞尔维亚语是一种南斯拉夫语，目前全球约有1200万人积极使用。它是被称为塞尔维亚-克罗地亚语的多中心语言的四种互通变体之一（其他变体包括克罗地亚语、波斯尼亚语和黑山语）。塞尔维亚-克罗地亚语在形态上十分丰富（Deli´c 等，2010年），包含大量词语的屈折变化，这是由于名词具有三种性、七个格，动词具有七种时态；这些屈折变化还涉及其他词性和词类，并且词语屈折中还会发生十二种语音变化（Klajn，2005）。塞尔维亚语也是欧洲唯一一种在形式上使用两种文字、其使用者在实际功能上也使用两种文字的语言，既使用西里尔字母，也使用拉丁字母（Magner，2001）。大多数塞尔维亚语使用者居住在塞尔维亚（根据2011年人口普查为6,330,919人[1]），但黑山、波斯尼亚和黑塞哥维那、克罗地亚、马其顿、斯洛文尼亚、阿尔巴尼亚、匈牙利、奥地利、瑞典、德国及其他国家也居住着数量可观的使用者。塞尔维亚语是塞尔维亚、波斯尼亚和黑塞哥维那、黑山的官方语言，同时在克罗地亚、马其顿、罗马尼亚、匈牙利、斯洛伐克和捷克共和国被承认为少数民族语言。塞尔维亚-克罗地亚语的变体（塞尔维亚语、克罗地亚语、波斯尼亚语和黑山语）约有1900万人使用，因此这些语言的重要性相当突出（Eberhard 等，2022）。

自然语言处理是人工智能的一个分支，研究如何分析和处理自然语言，并最终使计算机能够理解自然语言（Reshamwalaet al., 2013年）。因此，该领域致力于应对许多与人类语言/自然语言相关的挑战。尽管该领域的大部分工作主要针对英语开展，但也有针对其他语言的研究。

塞尔维亚语和塞尔维亚-克罗地亚语形态复杂、词形变化多样，而且可用资源相对匮乏，这给自然语言处理和语言技术带来了挑战。塞尔维亚-克罗地亚语丰富的形态特征使其尤为适合用来考察自然语言处理方法在词形变化多样的语言上的表现，以及如何在形态丰富且资源匮乏的语言中高效处理词形变化。从语言技术和自然语言处理的角度来看，不能孤立地看待塞尔维亚语，因为塞尔维亚语、克罗地亚语、波斯尼亚语和黑山语之间差异很小，而且针对其中一种变体开发的方法通常也能很好地适用于其他变体。尽管面临这些挑战，仍有若干倡议、组织和大量学术研究致力于应对塞尔维亚语的一些具体难题。研究人员开发了许多用于句法分析、分类和命名实体识别的资源与语料库，也提出了多种文档分析、分类、语义相似度乃至修辞格（如明喻）分析方法。

[1]https://data.stat.gov.rs/Home/Result/3102010401?languageCode=en-US

数字词汇资源的开发是每种语言一项重要且具有战略意义的任务，应当列为国家优先事项。自然语言处理的成果取决于可用数字资源的质量和数量，也取决于处理数字资源的工具是否可用且全面（Nenadi´c，2004）。我们的目标是收集塞尔维亚语文本数据处理方面现有的资源与方法，对其进行描述，并找出其中的不足，以便依据自然语言处理的发展趋势，利用最亟需的资源加以改进和扩充。据我们所知，这是首次对较新近、范围较广的塞尔维亚语自然语言处理资源与方法进行综述。

塞尔维亚自然语言处理资源与方法发展的简史

巴尔干地区首批数字语料库的开发工作，在世界首批数字语料库问世后不久便已启动；心理学家 Djordje Kosti´c 于 1957 年率先开展这项工作，旨在开发用于语音识别以及塞尔维亚-克罗地亚语机器翻译的语言技术。该语料库一直开发到 1962 年，但并未经过数字化处理，因此首个数字语料库于 1967 年在萨格勒布出版。这个语料库收录了由 ˇZeljko Bujas 整理的 Ivan Gunduli´c 史诗《Osman》。Dobri´c（2012年）介绍了 1950 至 1990 年间巴尔干西部地区语料库及语料库语言学的发展。此前已有研究综述了截至 2003 年主要在贝尔格莱德大学数学学院开发的语言资源与工具（Vitas 等，2003b，a）。2012 年，在名为 META-NET 的项目中，研究人员分析了欧盟 23 种官方语言的语言资源，并将《数字时代的塞尔维亚语》一书作为白皮书的一部分出版（Vitas 等，2012年）。区域语言数据倡议（ReLDI）项目为推动开放语言资源对塞尔维亚语及相关语言的重要性和价值作出了重要贡献（Samardˇzi´c 等，2015）。Batanovi´c 等人（2020年）简要介绍了在 ReLDI 项目中开发或独立构建、可开放免费获取的塞尔维亚语处理语言资源。

本文旨在回顾多年来为塞尔维亚语开发的语料库、资源、方法、模型和工具。我们有意将本次综述限定为塞尔维亚语。尽管我们承认，有些方法同样适用于塞尔维亚-克罗地亚语族中的相关语言，但这些语言之间仍存在细微差异，这会给资源和方法的评估与比较带来挑战。

2 综述方法

为了涵盖所有从事塞尔维亚语自然语言处理的作者，我们从塞尔维亚国家学位论文库（https://nardus.mpn.gov.rs/）开始。通过使用诸如自然语言之类的关键词

通过检索处理、自然语言处理、文本挖掘、文本数据处理、计算语言学、电子词典、语料库、情感分析、情绪分析、文本分类、词汇资源，以及其他同义词和相关术语，我们找到了包含这些关键词的学位论文。从中选出最相关的论文，即涉及塞尔维亚语自然语言处理的论文。此外，我们通过检索塞尔维亚各大学中已知的自然语言处理科学家、导师和研究团队，找出了一批学位论文。我们还通过检索担任导师或论文委员会成员的知名科学家，找到了相关学位论文。最终选出了自然语言处理领域的 29 篇学位论文。通过分析这些论文及其引用文献，我们找到了 316 篇与塞尔维亚语自然语言处理相关的论文。随后，我们在 Google Scholar 上针对知名作者（或作者团队）和选定主题进行了进一步检索。

我们审阅了找到的学位论文和论文，排除了与塞尔维亚语自然语言处理无关的内容，并根据主题和发表日期对其进行分类。本综述沿用这一分类方式，每个章节涵盖自然语言处理的一个广泛领域。各章节的内容主要按时间顺序编排。

3 语料库

语料库是一组机器可读文本，代表某种语言或文本类型的样本。语料库可以根据其媒介、范围（规模）、领域、用途、时期、来源、标注方法、所涉及的语言数量等参数进行分类（Vitas等，2003b）。鉴于语料库可以包含一种或多种语言的文本，因此可分为单语语料库和多语语料库。根据这一分类，我们将介绍塞尔维亚语语料库。

3.1 单语语料库

塞尔维亚-克罗地亚语历时语料库（DCSCL，表1）由Kosti´c教授于2003年进行数字化，收录了从12世纪到20世纪的文本，并分为五个时间样本。该语料库包含1100万个词，这些词均经过人工标注，附有词元以及各种形态类别的信息，例如性别、数、格、人物、时态等（Kosti´c，2014年）。1981年，数学学院的自然语言处理小组（NLP_MATF）启动了当代塞尔维亚语语料库的开发工作。该语料库的首个版本名为“当代塞尔维亚语未标注语料库”（UCCSL，表1），于2003年创建。该语料库收录了20世纪及之后出版的文学作品，且没有任何标注。随后，研究人员将书目信息纳入语料文本，对语料库进行了扩充；这一新版本名为*“SrpKor2003”*（SrpKor2003，表1）（Krstev和Vitas，2005年；Utvi´c，2014年）。

大多数单语语料库都带有形态句法标注，以及有关语料文本的书目信息。形态句法标注是一种语言学标注，它会为词元添加标签：词类（词性标注）、规范形式或词元（词形还原），以及形态词类。通过扩充 SrpKor2003，创建了当代塞尔维亚语语料库的新版本*"SrpKor2013"*（SrpKor2013，表1），其中包含20世纪和21世纪塞尔维亚作家的文学和艺术作品，以及科学文本、行政文本和通用文本。当代塞尔维亚语语料库包含书目数据，并已自动完成形态句法标注（包括词性和词元）。该语料库包含超过1.22亿个词。其子集"现代塞尔维亚语词形还原语料库"（SrpLemKor，表1）包含370万个语料词元。两个语料库均可在注册后依据许可证获取（Popovi´c，2010年；Utvi´c，2011）。

在贝尔格莱德大学数学学院现有的塞尔维亚语/塞尔维亚-克罗地亚语语料库中[2],，还包括以下单语语料库。*Henning* 的塞尔维亚-克罗地亚语语料库（HennC，表1）包含约70万词的塞尔维亚-克罗地亚语文本。这些文本取自现代南斯拉夫小说，涵盖所有使用塞尔维亚-克罗地亚语的地区（塞尔维亚、克罗地亚、黑山和波斯尼亚和黑塞哥维那）（语料库等，1992年）。*Vuk* 民间谚语未标注语料库（UnVukC，表1）收录了民间谚语及 Vuk 对其所作的评论（Krstev，1997年）。此外，Vuk 的明喻集还通过运用语法规则、机器学习和人工审查得到了扩充。最终，建立了一个包含852条明喻的当代明喻语料库（VukSimC，表1）[3]（Milosevi´c 和 Nenadi´c，2016年，2018）。*2000*年选举危机语料库收录了日报《政治报》从9月10日到10月20日的全部网络广播，而塞尔维亚语标注语料库包含带有最小结构标签集的文本；这两个语料库均缺少规模方面的详细信息，但可从同一来源获取[4]。

还有一些较小的语料库，主要针对特定领域（医学、法律等）和特定目的（命名实体识别、语义相似度等）收集。其中包括语料库 $MRCOR1$ 和 $MRCOR2$（表1），这些语料库由来自塞尔维亚32家医疗中心、经过审阅的医疗报告组成。主要数据集包含2212份诊断为麻疹的医疗报告。另一个数据集由2000年份包含十种不同类型诊断的医疗报告组成。医疗报告中的医学术语和非医学术语均由人工标注。每份医疗报告都被分配了一个诊断代码（Avdi´c等，2020年）。一个由贝尔格莱德大学牙科学院的100份出院清单和50份医生报告组成的语料库（DMRC，表1）被用于评估该系统自动分析医学叙事文本中时间表达的有效性。此前，这些文本

[2]http://www.korpus.matf.bg.ac.rs/prezentacija/korpusi.html [3] https://ezbirka.starisloveni.com [4] http://www.korpus.matf.bg.ac.rs/prezentacija/korpusi.html

Reference
(Kostić, 2014)
(Krstev and Vitas, 2005; Utvić, 2014)
(Krstev and Vitas, 2005; Utvić, 2014)
(Popović, 2010; Utvić, 2011)
(Popović, 2010; Utvić, 2011)
(Corpora etc, 1992)
(Krstev, 1997)
(Milosević and Nenadić, 2016, 2018)
(Avdić, 2021)
(Avdić, 2021)
(Jaćimović et al., 2015)
(Petrović and Stanković, 2019)
(Vasiljević, 2015)
(Sečujski and Delić, 2008)
(Krstev et al., 2012)
(Miličević and Ljubešić, 2016)
(Batanović et al., 2018; Batanović V et al., 2018)
(Batanović et al., 2011)
(Batanović et al., 2018)
(Stanković et al., 2021)
(Todorović et al., 2021)
(Stanković et al., 2020)
(Utvić et al., 2019)
(Vuković, 2020)
(Ljubešić et al., 2022)
(Lemmenmeier-Batinić et al., 2021)
(Bogetić and Batanović, 2020a)
(Bogetić and Batanović, 2020b)
(Ljubešić and Klubička, 2016; Ljubešić and Klubička, 2016)
(Bogdanović and Tošić, 2022)

ries, PoS - Part of Speech Tagging, S - structural,
cies, P - plausibility, MT - medical terms, NMT -
nnotated

已自动去除个人身份信息，但时间表达式未作更改（Ja´cimovi´c 等，2015）。*LAWC* 数据集（表 1）包含 1120 部法律文本，共切分为 59167 篇法律单独条文文本（Petrovi´c 和 Stankovi´c，2019）。语料库*LTC*（表 1）由电子版法律文本组成，这些文本可在塞尔维亚共和国国民议会网站上获取。截至 2014 年 5 月底通过的法律包含 681 篇文本（Vasiljevi´c，2015）。

从事自动语音识别（ASR）的 AlfaNum 建立了其资源*AlfaNum Text Corpus*（ATC，表 1）。该语料库按形态类别和重音进行标注，包含约 200,000 个词（Se˘cujski 和 Deli´c，2008年）。塞尔维亚语命名实体评测语料库（SrpNEval，表 1）由 2000 篇来自 2005 年和 2006 年塞尔维亚日报的简短新闻组成。该语料库采用塞尔维亚语的两种官方文字：西里尔字母和拉丁字母（Krstev 等，2012年）。*ReLDI-NormTagNER-sr 2.1*（NormTagNER，表 1）是一个人工标注的塞尔维亚语推文语料库，用于评估非标准塞尔维亚语的分词、句子切分、词语规范化、形态句法标注、词形还原和命名实体识别（Mili˘cevi´c 和 Ljube˘si´c，2016年）。*SETimes.SR*（SETimes.SR，表 1）是一个塞尔维亚语参考训练语料库，经过多层标注。SETimes.SR 中的文本取自多语种平行语料库*SETimes*（SETimes，表 2），该语料库收录了已停刊的东南欧时报新闻门户网站上的新闻文章（Batanovi´c 等，2018），（Batanovi´cV 等，2018）。塞尔维亚语释义语料库（paraphrase.sr，表 2）收集了来自网络新闻媒体的句子。人工为每对句子赋予二元相似度分数，以表示这对句子的语义是否足够相似、可视为近似释义（Batanovi´c 等，2011）。另一个用于确定语义相似度的语料库塞尔维亚语简短新闻文本语料库 -（STS.news.sr，表 2）由 1192 对从互联网新闻来源收集的塞尔维亚语句子组成（Batanovi´c 等，2018）。

19世纪40年代至20世纪20年代的塞尔维亚语旧小说收录于 *SrELTeC*（SrELTeC，表1）中，并作为 COST项目CA16204的一部分以数字形式保存（Stankovi´c 等，2021）。*ELTeC* 的塞尔维亚语部分包含120部小说（Odebrecht 等，2021）。这些小说带有结构性标注，并进行了句子切分；词语经过词性标注和词形还原，同时标注了七类命名实体。通过 ELG[5] 门户可获得的其他资源包括 *SrpELTeC-gold*、*SrpKor4Tagging*和*RudKorP*（表1）。用于训练命名实体识别的语料库SrpELTeC-gold 是塞尔维亚语文学语料库的一个子语料库，其中的命名实体由 SrpNER（Krstev等，2014）系统进行了标注（Todorovi´c等，2021）。SrpKor4Tagging 语料库由塞尔维亚语文学文本和行政-文本合并构成（Stankovi´c 等，2020）。RudKorP 语料库包含采矿和矿物原料加工领域的文本

[5]https://live.european-language-grid.eu/

这些材料由贝尔格莱德大学矿业与地质学院创建（Utvi´c et al., 2019）。通过 CLARIN.SI 平台还可以获取更多语料库，详见表 1 底部。*TorlakKor* 是一个语料库，收录了对蒂莫克（塞尔维亚东南部的一个地区）当地居民访谈的转录文本（Vukovi´c, 2020）。*COPA-SR* 数据集（塞尔维亚语合理选项选择）是英语 *COPA*数据集的译本（Ljubeˇsi´c et al., 2022）。*CorFoA* 是一个塞尔维亚语称呼形式语料库，收录了对 19 名参与者进行传记访谈的转录文本（Lemmenmeier-Batini´c et al., 2021）。*MLNews* 是一个综合性语料库，收录了与塞尔维亚语相关的新闻文章。该语料库还配有一个单独的语料库，收录市民对新闻文章的在线评论，称为*MLN-COM* （Bogeti´c and Batanovi´c,2020a,b）。塞尔维亚语网络语料库 *srWaC* 通过抓取塞尔维亚的 .rs 顶级域名于 2014 年 构建（ Ljubeˇsi´c and Klubiˇcka, 2016; Ljubeˇsi´c and Klubiˇcka, 2016）。*CorLeg* 是一个塞尔维亚共和国法律文本语料库，利用从官方网站 [7]收集的大量塞尔维亚法律文本创建（Bogdanovi´c and Toˇsi´c, 2022）。

3.2 多语语料库

多语语料库是一种特殊类型的语料库，包含以多种语言撰写的文本。平行语料库同时收录原文及其一种或多种其他语言的译文，并以特定方式呈现，使它们在文档、章节、段落、句子或词语层面的逻辑结构明确对应。表2展示了包含塞尔维亚语原文或译文的多语语料库。早期尝试构建多语语料库的项目之一，是创建柏拉图《理想国》的对齐语料库，其中包含包括塞尔维亚语在内的21种语言的译文。该语料库已在句子层面完成标注，并用于工具开发和自动对齐（Krstev and Vitas，2011）。名为 MULTEXT-East[8] 的语言资源和工具，用于从中东欧语言（中欧和东欧语言）的语言语料库中提取信息，是作为 Multext 项目的一部分创建的。小说《1984》收录于这一平行且按句子对齐的语料库《Multext-East语料库》（乔治·奥威尔的《1984》）（G.O.1984，表2）中，同时还收录了多种其他语言的译本。Krstev 和 Vitas（2011）创作了该小说的塞尔维亚语译本，并按照 MULTEXT-East格式进行了形态句法标注；此前，他们已为塞尔维亚语制定了相应规范。平行语料库 Verne80days（表2）包含儒勒·凡尔纳小说《八十天环游地球》的法语原文及17种译本。各语言文本均在子句层面进行了对齐（Vitas 等，2008）。塞尔维亚语-法语语料库（SrpFranKor，表2）由31篇在子句层面对齐的文本组成。

[6]https://www.clarin.si [7]
https://www.pravno-informacioni-sistem.rs/ [8]
http://nl.ijs.si/ME/

这些文本最初以法语写成，随后被翻译成塞尔维亚语，反之亦然，是塞尔维亚语中的第一个双语语料库（Vitas 和 Krstev，2006；Vitas 等，2006）。*ParCoLab*（表2）是一个可在线检索的平行语料库，由法语、塞尔维亚语、英语、西班牙语和奥克西唐语的句子对齐文本组成。这些语言中的每一种同时都是源语言和翻译语言（Balvet 等，2014年）。

塞尔维亚语-英语语料库 (SrpEngKo，表 2) 是第二个双语集合。它由英语源文本及其塞尔维亚语译文，以及反向的译文组成；此外，还收录了若干篇最初以法语写成的文学作品的英语和塞尔维亚语对齐译文（Krstev 和 Vitas，2011）。语料库 *SETimes* 基于发布在新闻网站 SETimes.com 上的文章。新闻以十种语言发布，其中包括保加利亚语、波斯尼亚语、希腊语、英语、克罗地亚语、马其顿语、罗马尼亚语、阿尔巴尼亚语和塞尔维亚语。SETimes 的一部分子语料库*BALKANTIMES* 用于扩展 SrpEngKo（Batanovi´c 等，2018）。法律、商业、教育和医疗保健领域的平行文本也被添加到 SrpEngKor 中，由此创建了子语料库 塞尔维亚语-英语法律、金融、教育与医疗保健 (SELFEH，表 2)。SELFEH 由近 150 篇平行文本组成，曾用于术语提取和机器翻译研究，也用于测试各种塞尔维亚语标注器（Utvi´c，2011）。另一个塞尔维亚语-英语语料库是 *srenWaC*（表 2），其中包含从 .rs顶级域名抓取的句子对齐平行文本（Ljubeˇsi´c 等，2016）。除了 SrpFranKo 和 SrpEngKo 双语语料库外，还创建了一个类似的德语语料库（SrpNemKor，表 2）。其中包含 48,004 对塞尔维亚语和德语文学文本译文，并按句子级别进行了对齐。研究人员将两种语言文本中的命名实体标注工具以及术语提取工具应用于准备好的平行语料库（Andonovski 等，2019；Andonovski，2020）。

此外，还有一些多语言平行语料库，其中部分列于表格末尾（表 2）。*OpenSubtitles* 是一个数据库，包含约 400 万条电影和电视节目字幕的句子级翻译，涉及 62 种以上不同语言（Tiedemann，2012年）。波斯尼亚语、克罗地亚语和塞尔维亚语网络语料库 （BsHrSrWaC，表 2）是顶级域名网络语料库。它们被用于创建一种区分相近语言的方法，该方法仅基于爬取数据进行单元语法语言建模（Ljubeˇsi´c 和 Klubiˇcka，2014年）。Twitter 用户数据集（Twitter-HBS，表 2）由推文及其语言标签（波斯尼亚语、克罗地亚语、黑山语或塞尔维亚语）组成。创建该语料库的主要目标，是在 Twitter 用户层面区分关系密切的语言（Ljubeˇsi´c 和 Rupnik，2022）。*PE2rr* 语料库包含来自多个领域的源语言文本，以及自动生成的多种形态丰富语言译文、这些文本的译后编辑版本，以及对译后编辑过程中所产生错误的标注。该语料库包含西班牙语、德语、塞尔维亚语文本，

Reference
(Vitas et al., 1998) (Krstev and Vitas, 2011)
(Vitas et al., 2008)
(Vitas and Krstev, 2006; Vitas et al., 2006) (Balvet et al., 2014) (Krstev and Vitas, 2011) (Batanović et al., 2018) (Utvić, 2011)
(Ljubešić et al., 2016) (Andonovski et al., 2019; Andonovski, 2020) (Tiedemann, 2012) (Ljubešić and Klubička, 2014) (Ljubešić and Rupnik, 2022) (Popović and Arčan, 2016) (Ljubešić and Lauc, 2021) (Ljubešić et al., 2021)

N - linguistic, PoS - Part of Speech Tagging, UK
ype: AA - automated annotation MA - manually

ge

斯洛文尼亚语和英语（Popovi´c 和 Arˇcan，2016年）。*BERTi´c-data* 文本集包含超过80亿个词元，主要为从网络爬取的波斯尼亚语、克罗地亚语、黑山语或塞尔维亚语文本。该语料库用于训练 BERTi´c Transformer 模型（Ljubeˇsi´c 和 Lauc，2021年）。波斯尼亚语、克罗地亚语、马其顿语、黑山语、塞尔维亚语、塞尔维亚-克罗地亚语和斯洛文尼亚语维基百科的转储收集于可比语料库*CLASSLA-Wikipedia* （CLASSLA-Wiki，表2）中。使用 classla 软件包 [9], 对所有语言在其可用的各个层面进行了语言学标注（Ljubeˇsi´c 等，2021年）。情感分析语料库在单独的章节中介绍。

4 语言资源

4.1 词典和术语表

电子词典是指用于文本处理的词典。它包含解决分词、形态以及部分句法和语义文本处理问题所需的宝贵信息（Vitas 和 Krstev，2009）。文本的自动处理始于对单个词语的分析，这些词语是所分析文本的基本单位。有时，单个词语可能并不是处理自然语言最恰当的基本单位。因此，词典分为两类：单词词典，由单个词语组成；多词词典，由多词单位组成（Andonovski，2020年）。

计算语言学国际实验室网络 RELEX（Laporte，2003）创建了一种构建电子形态词典的模型，该模型已被众多从事自然语言处理的组织采用。*Unitex* [10] 系统使用按照这一模型开发的电子形态词典。这些词典采用 DELA 格式（*LADL* 电子词典——文献自动化与语言学实验室）。为区分单词单位和多词单位，这种电子词典被组织为两个独立的子系统：单词单位词典（DELAS——简单形式，DELAF——屈折形式）和多词单位词典（DELAC——复合形式，DELACF——复合屈折形式）。

基于这些模型，贝尔格莱德大学语言技术小组构建了塞尔维亚语拉丁字母和西里尔字母电子形态词典（SrbMD）（Krstev，1997年；Vitas 等，2003b,a；Krstev 等，2006,2010）。据（Mladenovi´c，2016年）介绍，SrpMD 系统目前包含 148,000 个词元和 1,000 多个最终转换器，可生成 500 多万个 DELAF 词形分析结果。Leximir 工具（Stankovi´c 等，2011）用作词典管理系统。它是一款多用途工具，可支持计算语言学家开发、维护和利用电子词典。

[9]https://pypi.org/project/classla/ [10]
https://unitexgramlab.org/language-resources

该重音—形态词典由诺维萨德技术科学学院创建，包含超过400万条词条。它用于塞尔维亚语文本转语音和自动语音识别系统中的上下文分析（Seˇcujski 和 Deli´c，2008年）。

Ljubeˇsi´c 等人（2015）介绍了 $MWELex$，这是一个多语言词汇资源，收录从已解析语料库中提取的克罗地亚语、斯洛文尼亚语和塞尔维亚语多词表达（MWE）。$srMWELex$ 词典 v0.5 是在 PARSEME COST 行动的短期科学任务期间自动构建的。它收录了使用 DepMWEx 工具从$srWaC$ v1.0 网络语料库中提取的多词表达候选项。该词典包含 22 290 条词条和 3 273 369 个多词单位。文献（Ljubeˇsi´c 等，2016年）介绍了可免费获取的形态词典 srLex。它包含 105 359 个词位和 5 327 361 个（词元、词元、MSD）三元组。

Miletic（2017）介绍了塞尔维亚语形态句法电子词典的创建过程。该词典以塞尔维亚-克罗地亚语版维基词典、人工进行词性标注的语料库以及专业命题列表为基础。该词典包含 1 226 638 百万个词形，涵盖 117 445 个词元，对应总计 3 066 214 个不重复的三元组（词形、词元、MSD——形态句法描述），旨在用于词性标注和句法分析任务。

DELAS-TOP 和 DELAS-PERS 是分别收录地理名称和塞尔维亚语人名的词典（Krstev 等，2008年；Pavlovi´c-Laˇzeti´c 等，2004年；Grass 等，2002年）。地理名称词典 DELA-TOP 收录了高中地图集级别的地理概念（约 20.000 个地名、山名和水名及其相应的派生形式）。人名词典根据 1993 年统计的贝尔格莱德 170 万居民姓名名单编制而成。以此名单为基础，编制了两部词典：收录名字的 DELA-FName 和收录姓氏的 DELA-LName（Vitas 等，2003a）。

图书馆学与信息科学词典包含塞尔维亚语、英语和德语中的术语（Kovaˇcevi´c等，2004），这些术语用于图书馆学、信息科学及相关领域的理论与实践。该词典的在线版本目前包含：40,000 条定义（其中约有 14,000条为塞尔维亚语）；900 条作为图书馆标准组成部分的术语定义或标注；2,300个国际和国家组织及机构的首字母缩略词；190 个相关网站的地址[11]。

电子地质词典（GeolISSTerm）是一套专门编制的基本地质概念和术语分类体系，用于满足信息技术需求，作为塞尔维亚地质信息系统（GeolISS）中领域构建的基础资源（Stankovi´c et al., 2011）。

（Vujiˇci´c-Stankovi´c 等，2014年）在 SrpMD 中扩充了 636 条简单词词条和 612 条来自烹饪领域的多词表达（MWE）词条。

[11]http://rbi.nb.rs/srlat/dict.html

Grljevi´c（2016年）在其博士论文中提供了若干用于教育领域情感分析的词典（情感词、特定领域短语、否定关键词，以及从语料库中识别出的停用词）。否定信号、否定量词和语气助词强化词被加入情感词典中（Ljaji´c 和 Marovac，2019年）。同样，为进行情感分析，还创建了面向特定领域的停用词集（Mladenovi´c，2016年）。下文另设一章介绍情感分析，其中将进一步描述用于情感分析的情感词词典及其他词典。

Avdi´c 等人（2020年）创建了塞尔维亚语医学词典：诊断名称（7942条）、诊断代码（14194条）、诊断的拉丁字母名称（3794条）、疗法（2232种药物和1317支安瓿，共2255个不同术语）、麻疹诊断代码 B05 对应的症状（95条）、专科（41条），以及医学领域的缩写。同一研究还创建了非医学词典，包括医学领域的否定符号集、地点和姓名。

Ostrogonac 等人（2020年）创建了一个塞尔维亚语职业领域词汇表。该词汇表有两个版本，分别收录 4 万和 8 万个单词，以及 3 万个词元，并已纳入 Python 库 *nlpheart*。

塞尔维亚语停用词词典（SSW 词典）收录了 1241 个不同的塞尔维亚语停用词。该词典依据塞尔维亚语语法编制，并通过与现有的塞尔维亚语停用词集及克罗地亚语停用词集进行比较而完善。塞尔维亚语 SSW 词典收录了词语的不同形式。每个词语都附有一个词类标签。SSW 词典以 CSV 文件形式提供，文件名为 SSWdictionary.csv。该文件包含两列：单词和标签。标签说明词语的类别：助动词（V）、代词（PRON）、副词（ADV）、介词（PREP）、连词（CONJ）、感叹词（EXCL）、助词（PART）和缩写（ABBR）（Marovac 等人，2021年）。

SrHurtLex（Stankovi´c 等，2020年）是一个用于检测塞尔维亚语辱骂词语的词典。它基于词汇数据库 Leximirka、塞尔维亚语形态词典系统 SrpMD 以及《塞尔维亚语词典》（DS）（Vujani´c，2007）创建，收录了这些词典中标注为增义词、贬义词、侮辱性词语、粗俗词语等的多词表达。

4.2 本体

“本体”一词源自哲学，代表研究现存概念（事物类型）及其关系的科学（Vuji˘ci´c-Stankovi´c，2016年）。在计算机科学中，本体是一种描述概念、概念之间的关系以及现有约束的结构。其目的是在人与计算机之间，以及计算机彼此之间自动共享和重用知识。参与共享过程的双方都必须对可交换的信息具备一定程度的理解。

类的层级结构称为分类体系。通常，本体用于描述特定领域的术语及其之间的关系。

描述专名及其关系的语义网络是在 Prolex 项目期间开发的（Krstev 等，2007）。它由 2000年个专名组成，主要是国家及其首都的名称。

RudOnto 是在贝尔格莱德矿业与地质学院开发的术语资源，也是塞尔维亚语采矿术语的参考资源。它由术语信息系统管理，旨在为采矿工程的细分领域生成衍生术语资源，例如开采规划与管理、矿山安全或采矿设备管理（Stankovi´c 等，2011 年）。

Toma˘sevi´c（2018）开发了一个采矿领域本体$RuDokOnto$，用于收集、描述和系统化采矿项目文档，涵盖采矿项目生命周期的各个阶段，并以一种能够链接其他相关本体的方式进行组织。

RetFig 是一个语言学领域的描述性形式本体，用于描述塞尔维亚语中的修辞格，共描述了 98 种修辞格（Mladenovi´c 和 Mitrovi´c，2013年）。

4.3 词汇网络

在传统词典中，词汇概念按字母顺序排列，并且每个词汇概念的所有可能含义都有定义。在词网中，表达一个概念的所有词语被归为一个同义词集（synset - 同义词集）。塞尔维亚语*WordNet*（*SerWN*）（Krstev等，2004年；Koeva 等，2008年）是塞尔维亚语的词汇语义网络。其开发始于 BalkanNet 项目（Mladenovi´c等，2020年），并于 2004年完成，当时包含 8000个同义词集。此后，词网的开发继续进行，尤其是在生物学、生物医学、心理语言学和烹饪领域等方面。其结构基本上与 PWN（普林斯顿词网（Miller和 Fellbaum，2007））相同，并使用节点（同义词集）及其之间的关系进行组织。同义词集中的每个词都表示为一个字符数组或词项，后面跟着具体同义词集中具体词项的含义。由于一个词可能具有多种含义，因此它可以是多个同义词集的一部分。

根据 Koeva 等人（2008年），塞尔维亚语 *WordNet* 由 13612 个同义词集、23139 个词项、18210 种关系、314 个派生词和 83 个导出词组成。

Krstev 等人（2014年）为塞尔维亚语烹饪领域构建了一个本体，并用该领域的同义词集扩充了塞尔维亚语 WordNet。该本体用于确定食谱之间的相似度以及进行查询扩展。

作为一项词汇资源，$SerWN$ 已应用于多成员词汇单元研究（Krstev 等，2010年；Mladenovi´c 等，2014年）、文本分类（Pavlovic-Lazetic 和 Graovac，2010年）、多语种数字数据库检索（Stankovi´c 等，2012年）、修辞格识别（Mitrovi´c 等，2017年；Mladenovi´c 和 Mitrovi´c，2013年；Mitrovic，2014年）、文本情感分析（Mitrovi´c 等，2015）等领域。

Vujicic-Stankovic 创建了一个烹饪领域本体，并从烹饪领域扩展了 SrWN，增加了 1,404 个同义词集，使其总共包含 1,797 个此类同义词集（Vujiˇci´c-Stankovi´c 等，2014年；Vujiˇci´c-Stankovi´c，2016年）。

Universal Dependencies (UD) 项目[12] 旨在为多种语言开发跨语言一致的树库标注，提供通用的类别清单和指南，以促进跨语言对相似结构进行一致的标注，同时在必要时允许进行语言特定的扩展。作为该项目的一部分，塞尔维亚语树库基于 SETimes语料库创建（Samardˇzi´c 等，2017）。

5 词汇和句法分析方法

5.1 转写和变音符号恢复

Ljubešić 等人（2016年）介绍了一种工具，可自动为可能缺少变音符号的文本恢复变音符号（例如，在必要时将“kuca”（狗）转换为“ku´ca”（房子））。该工具在标准语言上的准确率为 99.5%，在非标准语言上的准确率为 99.2%。

由于每个音都对应一个字符，塞尔维亚语支持转写。字符几乎可以直接从西里尔字母映射到拉丁字母，但有几个字母例外：它们会从单个西里尔字符映射为两个拉丁字符（例如 њ -> nj、љ -> lj 或 ђ -> dj）。塞尔维亚语西里尔字母与拉丁字母之间的转写系统自20世纪50年代起就已存在（Matthews，1952；Aurousseau，1953；Gerych，1965）。用于解决这一问题的较新工具包括 Python软件包 *nlpheart* （Ostrogonac 等，2020），它能够在西里尔字母和拉丁字母之间进行转换。

5.2 分词与词干提取

句子分词是将文本划分为句子的过程。词语分词旨在将句子划分为简单的单位，即词元，通常包括词语、数字和标点符号。存在多种能够对塞尔维亚语文本进行分词的多语言分词器。这些工具大多以 Python 模块的形式提供，例如 Cutter（Gra¨en 等，2018）、Spacy[13]，CLASSLA 和 Reldi[14] 分词器。Cutter 分词器有一个用于在线分词的变体[15]。CLASSLA 分词器是经过改进、适用于特定语言的 Stanford NLP Python 库——用于处理斯洛文尼亚语、克罗地亚语、塞尔维亚语、马其顿语和保加利亚语的 Stanza 分支）[16]。塞尔维亚语的 Turanjin 分词器以 PHP库的形式提供[17]。目前没有关于塞尔维亚语文档分词准确率的准确信息或比较。

[12]https://universaldependencies.org/introduction.html [13] https://spacy.io/api/tokenizer [14]https://github.com/clarinsi/reldi-tokeniser [15] https://pub.cl.uzh.ch/projects/sparcling/cutter/current/ [16] https://pypi.org/project/classla/ [17] https://github.com/turanjanin/serbian-language-tools

Ostrogonac等人（2020年）推出了用于塞尔维亚语文本处理的Python软件包 *nlpheart*，其中包括转写、分词、规范化，以及为应用机器学习模型进行自动预处理。

词干提取是去除单词词尾字母（如派生后缀）的过程。剩余部分是单词的简化形式，称为词干。词干不同于单词的词典形式（词元）。Keˇselj和ˇSipka（2008年）介绍了首个用于塞尔维亚语词干提取的工具（下文简称词干提取器），该工具基于规则（1000条规则），准确率为79%。Miloˇsevi´c（2012b）基于这一词干提取器创建了新的词干提取器，将规则数量减少至180条，准确率达到90%。文献中还有另一种方案，由S. Petkovi´c等人[18] 创建，该方案基于克罗地亚语词干提取器（针对克罗地亚语的精确率为0.986，召回率为0.961，F1值为0.973）[19]。关于该工具的词干提取准确率，目前没有相关信息。Batanovi´c等人将Keˇselj和ˇSipka（2008年）的最优词干提取器和贪心词干提取器重新实现，改进了Miloˇsevi´c（2012b）提出的贪心算法，并将Ljubeˇsi´c和Pandˇzi´c开发的克罗地亚语词干提取器重新实现为WEKA软件包（Holmes等人，1994年）——即Batanovi´c等人（2016年）中的SCStemmers。该工具是对Ljubeˇsi´c等人（2007）所述算法的改进。

词干不是词典中的词形，而是具有相同语义的词语中最常见的部分。因此，在某些规范化方法中，会使用 n-gram分析作为词干提取器的替代方法。这意味着，一个词可以被规范化为其字母的一个单独子字符串，其长度为n（tri-gram、tetra-gram 等）。其原因在于，n-gram 分析方法与语言无关，也就是说，它不需要任何规则、词典或语料库。Marovac 等人（2012年）在塞尔维亚语文本的规范化中使用了 n-gram 分析。

5.3 词形还原和词性标注

词形还原旨在确定单词的基本形态形式（词元），即词典中的词条。在文本挖掘中，这一步对于塞尔维亚语等屈折形态丰富的语言尤为重要。一个词可能对应多个词元，具体取决于语境，因此一些词形还原器会利用词性标注或形态句法描述标注所提供的信息，以提高准确率。

词形还原的方法有多种：基于规则的方法、简单的基于统计的方法，以及基于机器学习的方法（Akhmetov 等人，2020年）。

LemmaGen（Jurˇsic 等人，2010年）是一种学习算法，用于以改进的 RDR（逐步修正规则）树结构的形式自动生成词形还原规则。它与 CST（Dalianis 和 Jongejan，2006）进行了比较。

[18]Stefan Petkovi´c 和 Dragan Ivanovi´c，《塞尔维亚语词干提取器》，2019年。https://snowballstem.org/algorithms/serbian/stemmer.html（访问于 2022年4月26日）[19]Ljubeˇsi´c, Nikola. Pandˇzi´c, Ivan.《克罗地亚语词干提取器》，http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/

以及RDR（Plisson 等，2008年）词形还原算法及其在塞尔维亚语语料库上的词形还原准确率见第 3 节。

BTagger[20] （Gesmundo和 Samardzic，2012年）是一种双向标注器-词形还原器工具，实现了将词形还原视为标注的范式。模型在塞尔维亚语 G.O.1984语料库上进行训练，词形还原的总体准确率达到 97.72%，形态句法描述标注的总体准确率达到 86.65%。

Agi´c 等人（2013年）在塞尔维亚语语料库上测试了隐马尔可夫模型三元标注器 HunPos、具备词形还原功能的 PurePos、TreeTagger、支持向量机标注器 SVMTool 、 CST 数据驱动规则式词形还原器以及 BTagger，结果见表 3。

表 3 规范化和词性标注工具

工具标签	应用程序	语料库	准确率
Keselj词干提取器（Ke˘selj 和 ˘Sipka，2008年）	词干提取器	未知	79.0%
Milosevic词干提取器 (Milo˘sevi´c, 2012b)	词干提取器	Politika	90.0%
LemmaGen（Jur˘sic 等人，2010年）	词形还原器	Multext-East	最高可达 86.1%+-0.61%
CST（Jur˘sic 等人，2010年；Dalianis 和 Jongejan，2006）	词形还原器	Multext-East	64.0 %+-0.82%
RDR（Jur˘sic 等人，2010年；Plisson 等人，2008年）	词形还原器	Multext-East	63.8%+-0.80%
BTagger（Gesmundo 和 Samardzic，2012年）	词形还原器	G.O.1984	97.73%
PurePOS（Agi´c 等，2013年）	词形还原器	SETimes	86.63%
TnT 标注器（Gesmundo 和 Samardzic，2012年）	词性标注器	G.O.1984	85.47%
BTagger（Gesmundo 和 Samardzic，2012年）	词性标注器	G.O.1984	86.65%
HunPOS（Agi´c 等，2013年）	词性标注器	SETimes	95.47%
条件随机场（Ljube˘si´c 等，2016年）	词性标注器	500k	97.86%
HunPOS（Agi´c 等，2013年）	MSD标注器	SETimes	87.11% (+词汇 84.81%)
PurePOS（Agi´c 等，2013年）	MSD标注器	SETimes	74.4%
SVMTool（Agi´c 等，2013年）	MSD标注器	SETimes	84.99%
条件随机场（Ljube˘si´c 等，2016年）	MSD标注器	500k	92.33%
Reldi标注器（Ljube˘si´c and Dobrovoljc，2019年）	MSD标注器	CLARIN.SI	92.03%
StanfordNLP（Ljube˘si´c 和 Dobrovoljc，2019年）	MSD标注器	CLARIN.SI	95.23%

词性标注是一项自然语言处理任务，即为文本中的词语标注相应的语法类别（词性：动词、名词、形容词、代词等）。提供更精确语法类别信息的词性标注称为形态句法描述标注（形态句法标注——使用形态句法描述进行标注）。

关于用于词汇和句法分析、并考虑形态句法标签的有限状态自动机，参见（Krstev, 1997年）。

Se˘cujski和Kupusinac（Se˘cujski和Kupusinac）使用隐马尔可夫模型对Alfanum和G.O.1984语料库进行形态句法标注。标注的准确率在很大程度上取决于文本类型，有些文本比其他文本更适合自动标注。AlphaNum语料库的错误率为18.44%，而“1984”的错误率高达26.97%。

Popovi´c（2008年、2010年）在三个语料库（《赫尔辛基笔记》第*15*期：少数民族与法律、塞尔维亚广播法和联合国开发计划署研讨会材料、G.O.1984）上评估了五种标注器（Tree Tagger、SVMTool、Brill基于规则的标注器、Trigrams’n’Tags和MXPOST）。TnT表现最佳，而Tree Tagger和SVMTool标注器在特定情况下表现更好。

基于 CRF（条件随机场）的塞尔维亚语和克罗地亚语词性标注器见（Ljubeši´c et al., 2016年）中的介绍。该标注器使用克罗地亚语人工标注语料库进行训练，并结合 hrLex/srLex词典进行使用。

[20]https://github.com/agesmundo/BTagger

每种语言。语料库中使用的形态句法标签集是根据修订后的 MULTEXT-East V5 克罗地亚语和塞尔维亚语标签集的说明创建的。塞尔维亚语词性标注的准确率为：形态句法描述标注为 92.33%，词性标注为 97.86%。

表 3 展示了用于塞尔维亚语语料库的分词、词干提取、词形还原、词性标注和形态句法描述标注工具及其准确率。

6 分类

文本分类是根据文本内容将文本数据归入预定义组别或类别的过程。文本分类通常使用监督式机器学习技术来完成。

Graovac（2014年）提出了两种根据文本内容进行文本分类的方法。第一种方法将文档表示为一个配置文件，其中包含文档中出现的固定数量的字节 n 元组，并使用差异度量来确定文档所属的类别。这种方法与语言无关，无需对文本进行任何预处理，也不需要预先了解文本内容或文本所用的语言。第二种方法则利用塞尔维亚语词网和塞尔维亚语电子词典中包含的信息。

Petrovi´c 提出利用模型和神经网络，作为满足以下需求的一种潜在解决方案：对新颁布的法律及其他法规文本中的链接或引用进行机器预测（Petrovi´cand Jani˘cijevi´c，2019年；Petrovi´c，2020年）。神经网络（RNN——循环神经网络、CNN——卷积神经网络，以及 HAN——分层注意力网络模型）的训练和验证是在一个标注数据集上进行的。该数据集通过为法律文本的每个片段（每一条法律条文）分配相应的标签而构建，用以标明该片段中是否存在链接或引用。随后，训练过程基于一个大型数据集，其中包含 1120 篇法律文本，这些文本共分割为 59 167 条独立的法律条文。

对于所有方法，训练参数的数量均减少了 99% 以上。

相似度

Marovac 等人（2013年）提出了一种用于塞尔维亚语文档相似性搜索的方法。搜索查询和待搜索的文档都以词向量表示。文档分组采用 k 均值聚类算法，并利用词频和逆文档频率特征以及 n元语法提取关键词。查询与文档之间的相似度值通过余弦度量、杰卡德系数或欧几里得距离计算。Furlan 等人（2013年）提出了一种名为 LInSTSS 的新算法，该算法在确定两段短文本的语义相似度时，也会考虑这些文本所包含词语的具体特点。评估使用了同一项研究中创建的塞尔维亚语释义语料库。

Nikoli´c（2016年）使用“Apache Lucene”工具描述了电子政务中的一种相似性搜索解决方案。Petrovi´c 和 Stankovi´c（2019年）展示了不同的准备方法如何影响文本相似度的计算。

Batanovi´c（2020年）介绍了使用统计建模和机器学习处理语义任务的过程。STS.news.sr 是一个新闻语料库，为语义相似度任务创建并用于该任务，新闻之间的相似度以分数标注。该方法通过 STSFineGrain（Java）库实现，可在 GitHub 上获取。语义相似度采用了词对齐与词向量平均值相结合的方法。srWaC 语料库（塞尔维亚语网络语料库）用于创建词向量。研究评估了三种不同词干提取技术对塞尔维亚语文本相似度的影响。此外，还提出了一种新的相似度计算技术，称为词性与词频加权的短文本语义相似度。

6.2 情感分析

情感分析是分析并提取人们对文本某部分中提及的各种主题、产品和服务所表达的意见、想法和印象的过程。情感分析可以从多个层面进行研究：文档级、句子级、短语级和方面级（Wankhade 等，2022）。进行情感分析需要特定的词汇资源，例如情感词典、否定处理工具、修辞手法处理工具等。塞尔维亚语句子级情感分析的首批工具之一由 Miloˇsevi´c（2012a）提出。研究使用朴素贝叶斯（NB）算法对负面和正面情感进行二元分类。预处理过程中使用了专为此目的设计的词干提取器（Miloˇsevi´c，2012b）。系统会移除停用词，并通过在否定信号词（如 no、none）之后的词前加上“NE_”来处理否定。情感分析器被开发为网络工具，并向公众开放[21]。

研究使用最大熵（ME）、支持向量机（SVM）和朴素贝叶斯（NB）机器学习方法分析推文情感（Joli´c，2015）。文中提出了降低这些消息噪声的流程，以提高准确率。使用单元语法时，ME 方法的准确率最高，达到 80.5 %；而在采用单元语法和二元语法，并考虑否定表达和短语后，准确率提高至 82.7%。

Mladenovi´c 等人（2016年）选择了一种混合方法：使用通过形态词典 SrbMD 扩展了形态形式的情感极性词典，并使用塞尔维亚语WordNet添加同义词，以减少在形态丰富的语言中使用词干提取器所带来的缺点。一个情感极性词典被创建出来（Mladenovi´c，2016年），其中包含 1053个表达（以及 10704 个屈折形式），并将其分为 24 个情绪类别，同时还通过

[21]https://inspiratron.org/SerbianSentiment.php

同义词和短语。SentiWordNet 已与塞尔维亚语词网整合，为塞尔维亚语词网中的同义词集添加情感标签，共标注了 4044 个同义词集。研究人员还利用这些同义词集创建了另一本包含 971 种屈折形式的情感词典。采用词频-逆文档频率方法，从当代塞尔维亚语 1.22 亿词语料库 SrpKor2013 中选取了 577 个最高频词（1428 种屈折形式），构建停用词列表。采用词频方法生成了一个面向特定领域的停用词集合，包含 1372 种屈折形式。该方法使用含两个主题——“坏消息”和“好消息”——的新闻集（TrN，表 4）进行训练；该新闻集按情感自动分类并保持平衡。测试使用了两个数据集：从 TrN 以外的来源收集的新闻集（TsN，表 4；该数据集不平衡），以及从网站收集并根据所附评分标注情感的电影评论集（TsMR，表 4；该数据集不平衡）。研究人员利用这些来源开发了塞尔维亚语文档级情感分析框架（SAFOS），该框架采用最大熵方法，并使用单元语法、二元语法和三元语法特征。他们通过留出测试集和 10 折交叉验证（CV）评估 SAFOS 系统。经“情感特征”映射筛选的单元语法与二元语法特征组合，在两个留出测试集上都取得了最佳分类准确率（TsMR 集准确率为 78.3%，TsN 集准确率为 79.2%）。由于训练数据和测试数据来自同一领域，该系统在 10 折交叉验证中的表现更好，准确率达到 95.6%。

Grljevi´c（2016年）对社交网络内容进行了情感分析，以改善高等教育机构的业务。情感分析在两个粒度层级上进行：文档级和句子级。在教授和讲座在线评论集（ORPL，表4）上，研究人员既采用了基于规则的策略（基于为该领域需求手动构建且可用的词汇），也采用了基于机器学习算法（NB、SVM 和 k-Nearest Neighbor KNN）的方法。在使用机器学习算法进行情感分类时，SVM 算法表现最佳：在评论层面的准确率为84.94%，在句子层面的准确率为80.13%。情感分类使用情感词典完成，分别为1266个正面情感词和1521个负面情感词、强化词（95个）、中和词、否定词（31个）、领域专用短语（41个）、停用词（179个）以及其他会改变句中下一个词情感的词语引入了独立词典。评论层面的分类准确率为80.71%，句子层面的分类准确率为73.70%。

首个塞尔维亚语均衡且主题统一的情感分析数据集（SerbMR，表 4）由 Batanovi´c 等人于 2016 年构建，并在线提供两种版本：一种包含两类情感极性（正面和负面；1682 篇文档），另一种包含三类情感极性（正面、中性和负面；2523 篇文档）。该数据集中的情感标签，是通过自动转换每条评论所附的数值评分获得的，评分由

其作者。对该数据集进行了研究，以确定用于情感分类的最佳机器学习特征和简单文本处理选项。通过将获得的最优属性与NBSVM（多项式朴素贝叶斯分类器和支持向量方法分类器的组合）相结合，他们在二分类中取得了最高85.55% 的准确率，在三分类中取得了最高62.69%的准确率。通过比较不同的形态归一化方法，得出结论：在情感分析中，使用词干提取器优于词形还原。Ljubeˇsi´c 和 Pandˇzi´c在数据集SerbMR 上取得了最佳准确率结果：二分类为86.11%，三分类最高为63.02%（Batanovi´c 和 Nikoli´c，2017 年）。

根据上述研究，仅识别否定的存在不足以确定情感。（Batanovi´c 等，2016）中的电影评论语料采用了传统的否定处理方法，即改变否定信号之后词语的极性。对于三分类任务，标记否定之后的两个词使情感分析准确率提升最为显著（0.94%）；对于二分类任务，只标记否定之后的第一个词则能带来最佳的准确率提升（0.66%）。Ljaji´c 和 Marovac（2019）描述了语义否定处理规则，这些规则提升了短篇非正式文本的情感分类效果。这些规则在一组涉及公众人物、并经过人工情感标注的推文上进行了测试（TWPP，表4）。一种基于所提否定处理规则、使用附加属性的机器学习方法，使三分类任务在一组推文上的情感分析准确率最高提升了 1.45%，二分类任务则最高提升了 0.82%。将此方法应用于包含否定的推文集时，三分类任务的情感分析准确率最高提升了 2.65%，二分类任务则最高提升了 1.65%。为实现本研究的目标，研究人员构建了否定信号词典（25 个词）、否定量词词典（56 个词）和程度强化词词典，以及一部包含 5632 个情感词的情感词典（经词形归一化后，包含 4058 个负面词和 1574 个正面词）。研究人员在这组推文上考察了不同词形归一化方法对情感分析的影响，发现使用词干提取器（Miloˇsevi´c，2012b）优于使用词形词典 SrbMD 进行归一化（准确率为 85.27%）；将词语缩减为 4-gram 则能以较少的资源消耗取得良好效果（Ljaji´c 等，2019）。

基于方面的情感分析旨在识别情感极性（负面、中性、正面），并确定句子中的方面（目标情感）。Nikoli´c 等人（2020年）提出了一种针对塞尔维亚语学生意见调查的基于方面的情感分析方法。研究使用了两组数据进行情感分析，并在文本最细粒度的层次上开展分析，即句子片段（短语和句子）层面。

官方学生调查（OSS，表4）合集构成了第一个数据集，而第二个数据集是此前为该论文创建的教授和讲师在线评论（OSPL，表4）（Grljevi´c，2016年）。The

OSS 和 OSPL 语料库先自动标注情感倾向（负面、中性、正面），再手动标注方面（从讲座、帮助程度、材料和组织等较低层级特征，到教授、课程及其他等较高层级方面）。方面分类采用级联分类器（一组经过训练、用于区分两个不同方面的 SVM 二元分类器）。方面分析的质量受语料库影响，OSS 语料库和 OSPL 语料库的 F 值分别为 0.89 和 0.78。

表 4 用于情感分析的语料库

语料库标签	文本类型	数量 of 条目	标注 [1]	参考文献
TrN	News	2000	S; AA	（Mladenovi´c 等，2016年）
TsN	News	779	S; AA	（Mladenovi´c 等，2016年）
TsMR	电影评论	2237	S; AA	（Mladenovi´c 等，2016年）
ORPL	教育评论	3863	S, A; AA + MA	(Grljevi´c，2016年)
OSS	教育评论	2472	S,A; AA +MA	(Nikoli´c 等，2020年)
SerbMR	News	2523	S;A A	（Batanovi´c 等，2016年）
SentiComments.SR	短文本	3490	S; MA	(Batanovi´c，2020年)
ParlaSent-BCS v1.0	句子 of 议会的辩论	2600	S; MA	（Mochtak 等人，2022）
TWPP	推文	7664	S; MA	(Ljaji´c 和 Marovac，2019年)
TWVA	推文	8817	S, R; MA	（Ljaji´c 等，2022）
MRSA	音乐评论	1830	S; AA	(Dra´skovi´c 等, 2022)
SMSSA	短信	6171	S; MA	(Sandrih，2019年)
TW15	推文	1643735	S; MA	(Mozeti˘c 等，2016年)

[1]标注目标：S - 情感，A - 方面，R - 相关性；标注类型：AA - 自动标注，MA - 手动标注

情感分析包括极性检测、主观性检测、讽刺检测等特定子任务。为满足处理特定任务的要求，并支持对情感进行多种解读，研究者设计了一种包含六种情感标签的标注方案（Batanovi´c 等，2020年）。SentiComments.SR（表4）是一个塞尔维亚语短文本语料库，采用这种多层级标注方案进行了人工标注。该语料库包含3490条简短的电影评论（长度不超过50个词元）（Batanovi´c，2020年）。研究人员在该语料库上评估了使用词袋和/或嵌入词袋特征的线性分类器，并考察了不同形态归一化方法及否定处理技术的影响。结合词袋与嵌入词袋属性后，所有情感分析子任务的分类效果均显著提升（F-测度：极性0.783，主观性0.885，四分类情感分析0.655，六分类情感分析0.586）。由于语料库中的讽刺文本数量不足，讽刺检测的结果不具代表性。

情感词典 Senti-Pol-sr（Stankovi´c 等，2022）基于三个现有词典（NRC、AFFIN 和 Bing）创建，并经过人工校正。该词典包含 6454 个不同的词元。其初始版本现已提供。

该词典被用于对从 SrpELTeC 提取的均衡数据集进行情感分析。该数据集包含 1089 个经人工标注的句子，正面、中性和负面情感极性类别各有 363 个实例。该方法取得了最佳

使用结合我们的情感词典特征的词袋方法，通过 MNB 在 SrpELTeC 二分类上取得了 87.8% 的准确率，在 SrpELTeC三分类上取得了71.9% 的准确率。使用 LR、NB、决策树、随机森林、SVN和 k-NN 方法训练的模型结果显示，LR 的准确率最高，达到87.8%。研究还表明，在带标签的电影评论数据集（SerbMR）上进行训练，无法将其成功用于老小说的句子情感分析。Draˇskovi´c 等人（2022）利用三种不同的数据集开发了一个情感分析机器学习模型。第一组数据（MRSA，表 4）是本研究通过收集13个门户网站上的音乐评论创建的，并确保了数据集的平衡。第二组数据集是前文提到的电影评论数据集，第三组则是音乐专辑评论数据集——MARD。MARD 最初由英语构成，随后使用 Google TranslateAPI翻译成塞尔维亚语。标准分类模型（NB、LR 和 SVM）以及混合模型（将线性模型与 NB 结合）被应用于这些数据集。混合模型 NB-LR 取得了平均较好的结果（三分类为 58%，二分类为 79%）。研究表明，将电影评论和音乐评论数据集合并使用，可以提高模型质量。由于词汇和评论写作风格不同，以及翻译文本的质量问题，将从英语翻译而来的评论纳入模型并不能改善性能。ˇSandrih（2019年）介绍了表情符号的影响、非正式语言、词汇特征以及短信数据集（SMSSA，表4）中与情绪相关的其他语言特征。他们选取了621个特征，并将其分为三大类：词汇特征（基于符号和词语）、句法特征（表情符号、缩写）和文体特征。使用线性SVM 分类后，准确率达到 92.3%。研究建议采用基于句子的情感分类以及情绪识别，以改进短信的分类。Mozetiˇc和 Grˇcar（Mozetiˇc和 Grˇcar）通过分析 15 种不同欧洲语言的160 万条人工标注推文（其中 73,783条推文为塞尔维亚语）（TW15，表 4），发现分类模型的质量在很大程度上取决于训练数据的质量和规模，而非训练模型的类型。根据所进行的实验，研究表明，顶尖分类模型的性能之间不存在统计学上的显著差异（其中五个模型基于 SVM，同时以 NB 分类器作为参考）。

迁移学习是人工智能领域的一项先进技术，它可以将预训练模型的知识迁移到新模型中。迁移学习常用于情感分析，以对情感极性进行分类；尤其是在缺乏大型标注数据集的情况下，它也能取得良好效果。

Batanovic 展示了将基于 Transformer架构的神经语言模型应用于 SentiComments.SR 语料库短文本情感分析子任务的结果（Batanovi´c 等，2020年）。使用了三种基于 Transformer 的模型：多语言 BERT（Devlin 等，2018）、多语言 distilBERT（Sanh 等，2019）和XLMMLM（Conneau 和 Lample，2019）。对多语言基于Transformer的模型进行微调，在所有情感分析子任务中均获得了与线性模型相同或更好的性能。对于每个

子任务中，XLM MLM 的 F 值表现最佳：极性分类为 0.793，主观性分类为 0.887，四分类情感分析为 0.686，六分类情感分析为 0.627。

Mochtak 等人（2022年）基于议会讨论样本，证明使用 Transformer 模型的结果明显优于使用更简单架构所获得的结果。该数据集由前南斯拉夫地区——波斯尼亚和黑塞哥维那、克罗地亚及塞尔维亚——议会会议语料库中的平均长度句子组成。研究使用克罗地亚语黄金标准情感词典（Glavaˇs 等人，2012年）（通过基于规则的克罗地亚语—塞尔维亚语翻译器译为塞尔维亚语（Klubiˇcka 等人，2016年）），从中选取 2600 个句子（ParlaSent-BCS v1.0，表 4）构建数据集，其中 876 个仅含正面词语，876 个仅含负面词语，866 个不含情感词语。该数据集包含来自塞尔维亚议会的 1059 个句子。数据集采用 Batanovi´c 等人（2020年）所述的多层级标注方案进行人工标注，并可在线获取。研究在句子层面应用了情感分析方法，并比较了四种 Transformer 模型的分类结果：使用 CLARIN.SI 预训练词嵌入的 FastText（Bojanowski 等人，2017年）（Ljubeˇsi´c 和 Erjavec，2018）、XLM-R（Con-neau 等人，2019年）、CroSloEngual BERT（Ulˇcar 和 Robnik-ˇSikonja，2020年）以及 BERTi´c（Ljubeˇsi´c 和 Lauc，2021年）。与其他模型相比，BERTi´c 的表现最佳（模型宏平均 F1 为 0.7941 ± 0.0101）。与波斯尼亚语和克罗地亚语相比，塞尔维亚语被证明是最难预测的语言。Ljaji´c 等人（2022）使用 BERTi´c 进行情感分析，扩展了一个标注数据集；该数据集曾用于分析有关 COVID-19 疫苗接种、含有负面情绪的推文主题。一组 8817 条塞尔维亚语疫苗接种相关推文（TWVA，表 4）经过人工标注，标明其与 COVID-19 疫苗接种情绪相关或无关。相关推文还经过人工标注，标记为正面、负面或中性。在该数据集上，BERTi´c 以 94.7% 的准确率正确判断推文是否相关，并以 85.7% 的准确率正确地将相关推文分类为负面、正面或中性。随后，使用该分类器扩充标注集：在最初人工标注的 1770 条负面推文之外，又自动标记了 1516 条负面推文，形成用于主题分析的数据集。主题分析采用了潜在狄利克雷分配（LDA）和非负矩阵分解（NMF）方法。主题分析突出显示了可能导致疫苗怀疑的因素：担忧不良反应、疫苗效力、测试不足、不信任当局以及阴谋论。

7 命名实体识别

命名实体识别（命名实体识别）是一项旨在定位并将非结构化文本中提及的命名实体分类为人物等类别的任务

人名、组织、地点、医学代码、时间表达式、数量、货币值、百分比等。作为一项自然语言处理任务，命名实体识别（NER）历史颇为悠久，随着20世纪90年代中期的消息理解会议（Sekine，2004）而逐渐受到关注。然而，直到2010年代，塞尔维亚语的命名实体识别才得到较为充分的研究。

Vitas 和 Pavlovi´c-La˘zeti´c（2008年）开发了一套系统，结合形态分析和词汇分析，并使用词典（塞尔维亚语和英语转写名字，以及地理位置）来识别人名和地理实体。该系统使用电子词典以及基于转换器的规则或语法，对专名和地缘政治实体进行消歧（Krstev 等，2007；Vitas 和 Pavlovi´c-La˘zeti´c，2008年）。

Ljube˘si´c 等人（2013年）提出了首个基于机器学习和条件随机场的系统，用于识别克罗地亚语和斯洛文尼亚语中的人名、组织和地点；这两种语言与塞尔维亚语关系密切。他们使用一组经过标注的网络语料库和新闻语料库（SETimes、Vjesnik，以及作为学生项目开发的斯洛文尼亚语和克罗地亚语语料库（Filipi´c 等人，2012年））来训练该方法。在特征方面，他们使用了语言特征，以及从大型未标注单语语料库中计算出的分布相似性特征。实验表明，分布特征可使F1分数提高7至8分，而形态特征还可额外提高3至4分。然而，随着数据集规模扩大，形态句法特征和分布特征对命名实体识别的重要性会降低。他们已公开提供构建该命名实体识别系统所用的资源。

另一种方法基于先前使用转换器和词库编码规则的应用（Krstev，1997年），并在此基础上加以改进，用于识别人物姓名和地缘政治名称（Krstev 等，2014年）。该方法使用词典匹配词元和短语，并使用 Unitex（Paumier 等，2002年）中的递归转换网络（语法图）消除歧义（例如，考虑格、数、性一致等语法规则）。研究人员报告称，该系统更注重精确率而非召回率，精确率为 0.96，召回率为 0.88。

为比较多语言对齐文本（双语平行文本）中的命名实体识别方法，研究人员开发了一个名为 NERosetta 的系统（Krstev 等，2013年；Krstev C 等，2013年）。为展示该系统，他们使用了 7 组涉及 5 种语言（法语、英语、希腊语、塞尔维亚语、克罗地亚语）的双语平行文本，以及 5 种不同的命名实体识别系统（1 种用于塞尔维亚语（Krstev 等，2014年），1 种用于克罗地亚语（Ljube˘si´c 等，2013年），1 种用于英语（Stanford NER），2 种用于法语）。评估的实体包括人物、组织和地点；其中一些命名实体识别系统还提供时间、日期、货币、百分数等标注。演示应用可在网上获取 [22]。

研究采用词典方法，并加入基于转换器的语法（Krstev 等，2014年），以新闻文章中标注的人名为基础创建金标准数据集。随后，该数据集被用于训练基于机器学习的方法，即 Stanford NER 和 Spacy

[22]http://www.korpus.matf.bg.ac.rs/nerosetta/

（ˇSandrih 等，2019年）。评估结果表明，基于规则的方法表现最佳（按 F1 分数衡量），而 Stanford NER 的召回率最高。

Tanasijevi´c（2019年）开发了一个为文化遗产文档添加元数据标签的系统。为此，她开发了一套系统，用于识别年份、人名等实体，以及已标注文档的主题。

基于Transformer 的模型也被引入用于塞尔维亚语、克罗地亚语和斯洛文尼亚语中的多项任务，包括命名实体识别（Ljubeˇsi´c和Lauc，2021年）。该模型在塞尔维亚语、波斯尼亚语、克罗地亚语和斯洛文尼亚语的网络爬取文本上进行了预训练，这些文本包含80亿个词元；随后，该模型在若干公开可用的数据集上针对命名实体识别进行了微调，例如 SETimes.SR（Batanovi´c 等，2018）、新闻文章语料库，或 ReLDI-sr（Ljubeˇsi´c 等，2017）这一标注推文语料库。作为参考，作者将该模型与 CroSloEngual BERT （ Ulˇcar 和 Robnik-ˇSikonja，2020 年）以及多语言 BERT （Devlin等，2018）进行了比较，结果显示，特定语言的 BERT 模型显著优于多语言 BERT。

除塞尔维亚语通用领域外，医学命名实体识别方面也已开展了相当多的工作。此前介绍的一个通用领域系统经过改造，用于临床文本的医学去标识化（Ja´cimovi´c 等，2014年，2015）。该系统利用词汇表和转换器语法规则识别人员、日期、地理位置、组织和数字。作者报告的总体 F1分数为 0.94。另一方面，Puflovi´c 等（2016）创建了一个基于字符和词 n 元语法的模型。他们使用的数据集来自一家神经科诊所，其系统旨在识别疾病名称、药物名称、缩写，以及表示剂量、日期或时间和医疗治疗效果的数字。他们对 100 份文档进行了人工检查，报告的准确率介于 64% 至 90% 之间。Avdi´c 等（2020）提出了一种用于医学术语识别的数学模型，并提出了三种方法。第一种方法基于术语的词典匹配。第二种方法使用一个公式为训练集中包含的词语赋予标签，其中置信度的计算方式为：某个词在训练集中被赋予特定标签的实例数除以该词在训练集中的总出现次数。第三种方法是第二种方法的扩展，针对存在错误和缩写的术语加入了若干规则，以便更好地进行标注。标注的实体包括医学术语（症状、症状描述、诊断、生化分析、拉丁词、器官解剖名称、疗法及其他医学术语）和非医学术语（数字、否定符号及其他词语）。表现最佳的是第三种模型，其 F1分数为 0.937；而医学术语的最高 F1分数为 0.896。Kaplar 等（2022）提出了基于深度神经网络和多语言语言模型的方法。他们使用了来自诊所的人工标注语料库

塞尔维亚大学临床中心肾脏病学科（203份出院摘要，由2名计算机科学博士生根据2012年i2b2时间关系挑战赛标注方案进行标注）。他们创建了基于条件随机场（条件随机场）、多语言Transformer（多语言BERT和XLM RoBERTa）、长短期记忆（长短期记忆）循环神经网络的模型，以及这些模型的集成模型。条件随机场方法使用了文献中常见的手工设计特征（词语、词干、词形、前3个词、后3个词等）。对于长短期记忆模型，作者先使用gensim的word2vec模型，再将嵌入输入长短期记忆网络，随后接入条件随机场词元分类器。研究表明，基于条件随机场的模型精确率最高，而多语言Transformer模型的召回率最高。基于长短期记忆网络-条件随机场的模型F1分数最高。通过多数投票集成这些模型，取得了最佳性能（F1分数为0.892）。

8 语言模型

语言模型确定序列中的词概率。为了创建语言模型，人们提出了许多方法，从简单地计算较大文本语料库中的词语出现次数，到添加更多词汇和句法特征来学习词概率。这些早期语言模型纯粹是统计模型，而自2014年以来，我们见证了神经语言模型的大量涌现——即基于神经网络的语言模型。

语言模型是许多自然语言任务的先决条件。因此，许多分类（Graovac, 2014年）、情感分析（Milo˘sevi´c，2012年a；Joli´c，2015；Grljevi´c，2016年；Batanovi´c 和 Nikoli´c，2017；Ljaji´c和 Marovac，2019年）或命名实体识别（˘Sandrih等，2019年）研究使用了传统的 n-gram语言模型，有时还结合了词汇、形态- 学或句法特征。这些系统已在本综述中作过描述。

Ostrogonac（2018）在其博士论文中回顾并比较了截至 2018 年的塞尔维亚语语言模型。在这项工作中，他提出了第一个塞尔维亚语神经语言模型，该模型基于在塞尔维亚语形态标注文本语料库上训练的循环神经网络。此外，他创建了一个混合模型，使用词性和词元，并将词语序列与语料库中的 n 元语法或部分词元化序列进行匹配。这些模型与更为传统的 n 元语法模型进行了比较，用于纠正文本中的语义和语法错误。通过设定具有形态特征的语言模型与不具有形态特征的语言模型之间对数似然差异的阈值来检测错误。尽管设定阈值可能颇具挑战，但这项研究展示了专门训练的塞尔维亚语神经语言模型的潜在应用场景。

国际研究人员为创建多语言神经语言模型付出了大量努力。其中一些模型也包含塞尔维亚语，例如 FastText（Bojanowski 等，2017）、多语言 BERT

（Devlin 等，2018）、XLM-R（Conneau 等，2019）以及XLM MLM（Conneau 和 Lample，2019）。Batanovi´c在其博士论文（Batanovi´c，2020年）中，比较了若干用于情感分析和文本相似度任务的n元语言模型。他进一步将这些语言模型和方法与经过微调的多语言、基于 Transformer 的模型（多语言BERT base（Devlin等，2018）、DistilBERT Multilingual（Sanh 等，2019）以及 XLM MLM（Conneau 和 Lample，2019））进行了比较，结果表明，在所有情况下，Transformer模型的表现都优于所有基于 n 元模型的模型（包括那些包含大量形态、词汇和句法特征的模型）。

第一种，也是截至本文撰写时唯一一种专门针对塞尔维亚语、克罗地亚语、波斯尼亚语和黑山语训练的基于 Transformer的语言模型，是BERTi´c（Ljubeˇsi´c 和 Lauc，2021年）。BERTi´c 采用 ELECTRA 方法（Clark 等，2020年）训练 Transformer 模型。该方法包括训练一个较小的生成器模型和主要的判别器模型，任务是判别某个特定词语究竟是文本中的原始词语，还是由生成器模型生成的词语。该模型基于一个语料库进行训练，该语料库包含从网络上抓取的、由塞尔维亚语、克罗地亚语、波斯尼亚语和黑山语组成的80亿个词元。虽然此前已有一个基于 BERT 的克罗地亚语和斯洛文尼亚语模型，名为 CroSloBERT（Ulˇcar和Robnik-ˇSikonja，2020年），但BERTi´c在几乎所有任务上都超过了它（形态学标注、命名实体识别、社交媒体地理位置预测、常识因果推理任务）。这主要归因于规模更大的语料库，以及所采用的 ELECTRA 方法的计算效率。

9 结论与未来方向

塞尔维亚语自然语言处理研究有着悠久的传统，可追溯至20世纪90年代下半叶。在此期间，人们探索了文本词汇、形态、句法和语义处理的多种方法。过去十年里，塞尔维亚语自然语言处理领域的研究人员数量和相关研究成果显著增加。塞尔维亚的几所大学和研究机构建立了自然语言研究团队。

塞尔维亚语是一种屈折变化丰富的语言，因此自然语言处理中的许多挑战都具有塞尔维亚语的特异性，例如如何实现最有效的分词、如何在各种任务中处理屈折变化、如何处理否定等。尽管已有一些研究探讨这些挑战，但它们仍是尚待解决的研究问题。转写、变音符号恢复、分词、词干提取、词元化和词性标注等基础词汇与形态任务已得到较为充分的研究，已有许多方法被提出、评估和比较。部分分类任务，例如情感分析，也得到了广泛研究。2012年后，情感分析似乎受到了广泛关注。针对多种命名实体（例如专有名词、人名和地点）的命名实体识别也已有相关研究。此外，也有少数方法被提出用于生物医学命名实体识别。

另一方面，针对塞尔维亚语，仍有一些方法和任务没有得到充分研究。除情感分析外，许多分类任务尚未得到探索，相关语言资源也十分匮乏。正如前文所述，目前仅有针对一组基础命名实体的方法被提出。领域特定分类和命名实体识别方法仍然缺失。

语义网、本体和语义网络方面的方法在塞尔维亚语自然语言处理领域尚未得到广泛发展，因为只有少数论文涉及这一主题。网络领域最重要的研究工作集中于开发塞尔维亚语 WordNet，但它更像是一个形态和词汇网络，而不是可以被视为语义网络的资源。许多任务所需的语言资源仍然缺失。

虽然基于语言特定 BERT 的模型已经完成训练，但目前只有一项计划致力于创建此类语言模型。此外，句子嵌入或文档嵌入等资源和方法尚未开发。这些方法也将为摘要生成、问答、语言特定语义搜索或机器翻译方法的创建作出重要贡献。

目前，大语言模型层出不穷，例如 GPT-3（Brown 等，2020年）、Lambda（Thoppilan 等，2022）或 ChatGPT（Ouyang 等，2022）。虽然这些模型支持多种语言，并能够生成塞尔维亚语文本，但目前针对塞尔维亚语进行提示工程或微调这些语言模型的研究还不多。

10 致谢

本文部分得到塞尔维亚共和国教育、科学与技术发展部项目编号 III44007 的资助。

参考文献

Agi´c, ˇZ.、N.Ljubeˇsi´c和D.Merkler，2013年。克罗地亚语和塞尔维亚语的词形还原与形态句法标注。载于第*4*届两年一度的波罗的—斯拉夫自然语言处理国际研讨会论文集，第48–57页。

Akhmetov、I.、A. Pak、I. Ualiyeva 和 A. Gelbukh。2020年。使用机器学习分类器实现高度独立于语言的词形还原。*Com-putaci´on y Sistemas 24*(3): 1353–1364 。

Andonovski, J. 2020年。 构建塞尔维亚语—德语文学语料库过程中的开放数据网络和语言资源。 博士学位论文。

Andonovski, J.、B. ˇSandrih 和 O. Kitanovi´c。2019年。基于词对齐的双语词汇提取，以改进语料库检索。《*The Electronic Library*》。

30 塞尔维亚语自然语言处理资源与方法综述Aurousseau, M. 1953. 西里尔字母文字的转写。*Nature* 171: 940–940。Avdi´c, A. 2021. 智慧医疗服务的实现及其与智慧城市概念的整合。博士学位论文。Avdi´c, A., U. Marovac, 和 D. Jankovi´c. 2020. 塞尔维亚语医疗报告中术语的自动标注。土耳其电气工程与计算机科学期刊 *28*(6): 3285–3303。Balvet, A., D. Stosic, 和 A. Miletic 2014. TALC-SEF：一个经人工修订的塞尔维亚语、英语和法语文学词性标注语料库。载于*LREC 2014*。Batanovi´c, V., M. Cvetanovi´c, 和 B. Nikoli´c 2018. 塞尔维亚语细粒度语义文本相似度。载于第十一届国际语言资源与评估会议论文集（*LREC 2018*）。Batanovi´c, V., M. Cvetanovi´c, 和 B. Nikoli´c. 2020. 面向资源受限环境的短文本情感表达、标注与分析通用框架。*PLoS One 15*(11): e0242050。Batanovi´c, V., B. Furlan, 和 B. Nikoli´c. 2011. 用于确定塞尔维亚语短文本语义相似度的软件系统。第*19*届电信论坛论文集（*TELFOR 2011*）: 1249–1252。Batanovi´c, V., N. Ljubeˇsi´c, 和 T. Samardˇzi´c 2018. SETimes.sr——塞尔维亚语参考训练语料库。载于*2018*年语言技术与数字人文会议论文集（*JT-DH 2018*），第 11–17 页。Batanovi´c, V., N. Ljubeˇsi´c, T. Samardˇzi´c, 和 M.M. Petrovi´c. 2020. 塞尔维亚语处理的开放资源与技术。《自由软件与开放硬件的应用》论文集。Batanovi´c, V. 和 B. Nikoli´c. 2017. 塞尔维亚语文档的情感分类：形态规范化与词嵌入的影响。*Telfor Journal 9*(2): 104–109。Batanovi´c, V., B. Nikoli´c, 和 M. Milosavljevi´c 2016. 资源受限语言情感分析的可靠基线：塞尔维亚语电影评论数据集。载于第十届国际语言资源与评估会议论文集（*LREC'16*），第 2688–2696 页。Batanovi´c, V.V. 2020. 资源受限自然语言短文本处理中的语义问题解决方法。博士学位论文，贝尔格莱德大学电气工程学院。Batanovi´c V, V., N. Ljubeˇsi´c, T. Samardzi´c, 和 T. Erjavec. 2018. SETimes.sr 1.0 训练语料库。

1Bogdanovi´c, M. 和 J.To˘si´c. 2022. 塞尔维亚共和国法律文本语料库 1.0。斯洛文尼亚语资源库 CLARIN.SI。Bogeti´c, K.和 V. Batanovi´c. 2020 年a.塞尔维亚语相关新闻文章标注语料库 MetaLangNEWS-sr。斯洛文尼亚语资源库 CLARIN.SI。

Bogeti´c, K. 和 V. Batanovi´c. 2020年b. 塞尔维亚语相关新闻评论标注语料库 MetaLangNEWS-COMMENTS-sr. 斯洛文尼亚语资源库 CLARIN.SI.

Bojanowski, P.、E. Grave、A. Joulin 和 T. Mikolov。2017。利用子词信息丰富词向量。计算语言学协会会刊 5：135–146。

Brown, T., B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan， P. Dhariwal, A. Nee- lakantan， P. Shyam， G. Sastry， A. Askell， 等。2020年。语言模型是少样本学习者。神经信息处理系统进展33：1877–1901。

Clark, K.、M.T. Luong、Q.V. Le 和 C.D. Manning。2020年。ELECTRA：将预训练文本编码器用作判别器，而非生成器。*arXiv* 预印本 *arXiv:2003.10555* .

Conneau, A., K. Khandelwal, N. Goyal,V.Chaudhary, G. Wenzek,F.Guzm´an, E. Grave, M. Ott,L. Zettlemoyer, and V. Stoyanov. 2019年. 无监- 督的大规模跨语言表示学习. *arXiv* 预印本 *arXiv:1911.02116* .

Conneau, A. 和 G. Lample. 2019年. 跨语言语言模型预训练。神经信息处理系统进展32。Corpora etc, C. 1992年. 塞尔维亚-克罗地亚语文本语料库。牛津文本档案馆。Dalianis, H. 和 B. Jongejan 2006. 手工制定与机器学习得到的屈折规则：euroling-siteseeker 词干提取器和 cst 词形还原器。载于第五届国际语言资源与评估会议论文集（*LREC'06*）。Deli´c, V., M. Se˘cujski, N. Jakovljevi´c, M. Janev, R. Obradovi´c, 和 D. Pekar. 2010年. 塞尔维亚语及相关南斯拉夫语支语言的语音技术。语音识别进展：141–164。Devlin, J., M.W. Chang, K. Lee, 和 K. Toutanova. 2018. BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv* 预印本 *arXiv:1810.04805* 。

32 塞尔维亚语自然语言处理资源与方法综述Dobri´c, N. 2012. 西巴尔干地区当代语言语料库——历史、现状与未来。《斯拉夫学评论》*60*(4): 677–692 . Dra˘skovi´c, D., D. Ze˘cevi´c, and B. Nikoli´c. 2022. 塞尔维亚语机器情感分析多语言模型的开发。《数学》*10*(18): 3236 . Eberhard, D.M., F.S. Gary, and D.F. Charles. 2022. Ethnologue：世界语言。 Filipi´c, L., T. Juri´c, and M. Stupar. 2012. 用机器识别克罗地亚语文本中的名称。 学生科学论文，校长奖，萨格勒布大学哲学学院 . Furlan, B., V. Batanovi´c, and B. Nikoli´c. 2013. 自然语言处理支持不足的语言中的短文本语义相似度。《决策支持系统》*55*(3): 710–719 . Gerych, I. 1965. 西里尔字母的转写. 博士学位论文，渥太华大学（加拿大）。 Gesmundo, A. and T. Samardzic 2012. 将塞尔维亚语词形还原作为类别标注与双向序列分类。载于 第八届国际语言资源与评估会议（*LREC'12*）论文集，第 2103–2106 页。 ˇGlava˘s, G., J. Snajder, and B. Dalbelo Ba˘si´c 2012. 克罗地亚语情感词典的半监督获取。载于 国际文本、语音与对话会议，第 166–173 页。Springer。 Gra¨en, J., M. Bertamini, M. Volk, M. Cieliebak, D. Tuggener, and F. Ben- ites 2018. Cutter——通用多语言分词器。载于 *CEUR*研讨会论文集，第 2226 号，第 75–81 页。CEUR-WS。 Graovac, J.B. 2014. 文本分类方法研究：数学模型与应用. 博士学位论文。 Grass, T., D. Maurel, and O. Piton 2002. 专名多语言数据库的描述。载于 自然语言处理进展：第三届国际会议，*PorTAL 2002*，葡萄牙法鲁，*2002*年*6*月*23–26*日，论文集，第 137–140 页。Springer。 Grljevi´c, O. 2016. 社交媒体内容中的情感分析：促进高等教育机构发展的工具. 博士学位论文。 Holmes, G., A. Donkin, and I.H. Witten 1994. Weka：机器学习工作台。载于*ANZIIS'94*——澳大利亚新西兰智能系统会议论文集

信息系统会议，第 357–361 页。IEEE。

Ja´cimovi´c, J., C. Krstev, 和 D. Jelovac，2014年。受保护健康信息的自动去标识化。载于 第*17*届国际多学科会议*“*信息社会*—IS 2014”*论文集，语言技术，*2014*年*10*月*9*日至*10*日，斯洛文尼亚卢布尔雅那，第73–78页。约瑟夫·斯特凡研究所，斯洛文尼亚卢布尔雅那。

Ja´cimovi´c, J.、C. Krstev 和 D. Jelovac。2015年。基于规则的医学叙事文本自动去标识系统。*Informatica 39*(1)。Joli´c, N. 2015。利用远程监督对 Twitter 帖子进行情感分类。Jurˇsic, M.、I. Mozetic、T. Erjavec 和 N. Lavrac。2010年。LemmaGen：使用归纳的涟漪式规则进行多语言词形还原。*Journal of UniversalComputer Science 16*(9): 1190–1214。Kaplar, A.、M. Stoˇsovi´c、A. Kaplar、V. Brkovi´c、R. Naumovi´c 和 A. Kovaˇcevi´c。2022。塞尔维亚电子健康记录中临床命名实体识别方法的评估。*International Journal of MedicalInformatics*: 104805。Keˇselj, V. 和 D. ˇSipka。2008年。一种基于后缀包含关系的方法，用于为资源稀缺的高度屈折语言构建词干提取器和词形还原器。*INFOtheca-Journal of Informatics & Librarianship* 9。Klajn, I. 2005。塞尔维亚语语法 （第1版）。贝尔格莱德：教科书与教学资源出版社。Klubiˇcka, F.、G. Ramırez-S´anchez 和 N. Ljubeˇsi´c。2016。克罗地亚语与塞尔维亚语之间基于规则的机器翻译器的协作开发。载于欧洲机器翻译协会第*19*届年会论文集，第361–367页。Koeva, S.、C. Krstev 和 D. Vitas。2008。词网中的形态语义关系——以两种斯拉夫语为例。载于全球词网会议，第239–253页。塞格德大学信息学系。Kosti´c, A. 2014。12至18世纪塞尔维亚语电子语料库。国家数字化中心评论 。Kovaˇcevi´c, L.、V. Injac 和 D. Begeniˇsi´c。2004。图书馆术语词典：英塞、塞英。塞尔维亚国家图书馆。Krstev, C. 1997。文本的信息学建模及其转换算法的一种方法。博士学位论文。

34 塞尔维亚语自然语言处理资源与方法综述Krstev, C., J. Ja´cimovi´c, and D. Vitas 2012. 塞尔维亚语中若干类命名实体的识别与规范化。载于第五届巴尔干信息学会议论文集，第 52–57 页。Krstev, C., D. Maurel, et al. 2007. 关于 Prolex 项目中专名语义和形态特性的札记。*LingvisticaeInvestigationes 30*(1): 115–133。Krstev, C., I. Obradovi´c, M. Utvi´c, and D. Vitas. 2014. 一种基于局部语法的命名实体识别系统。*Journal of Logic andComputation 24*(2): 473–489。Krstev, C., G. Pavlovi´c-La˘zeti´c, and I. Obradovi´c. 2004. 在开发塞尔维亚语词网时使用文本和词汇资源。*Romanian Journal ofInformation Science and Technology 7*(1-2): 147–161。Krstev, C., R. Stankovi´c, I. Obradovi´c, D. Vitas, and M. Utvi´c 2010. 多词单位形态词典的自动构建。载于国际自然语言处理会议，第 226–237 页。Springer。Krstev, C., S.V. Stankovi´c, and D. Vitas 2014. 烹饪领域中的近似度量：本体与词汇资源。载于第九届语言技术会议 *IS-LT* 论文集，第 38–43 页。Krstev, C. and D. Vitas 2005. 语料库与词典——相互不完备。载于语料库语言学会议论文集，第 14 卷，第 17 页。Krstev, C. and D. Vitas. 2011. 英塞对齐语料库。*ELL- SIIR* 论文集（英语语言与文学研究：形象、身份、现实）1: 495–508。Krstev, C., D. Vitas, and G. Pavlovi´c-La˘zeti´c 2008. 塞尔维亚-克罗地亚语形态句法处理的资源与方法。载于斯拉夫语言形式描述：第五届会议，第 3–17 页。Krstev, C., D. Vitas, and A. Savary 2006. 构建全面的塞尔维亚语复合词词典的先决条件。载于国际自然语言处理会议（芬兰），第 552–563 页。Springer。Krstev, C., A. Ze˘cevi´c, D. Vitas, and T. Kyriakopoulou 2013. NERosetta——命名实体标注简介。载于第六届语言与技术会议，第 168–172 页。Krstev C, C., A. Ze˘cevi´c, D. Vitas, and T. Kyriacopoulou 2013. 面向命名实体多语言空间的 NERosetta。载于语言与技术会议，第 327–340 页。Springer。

5 Laporte, E. 2003. RELEX 网络。获取自 *http://infolingu. univmlv. fr/Relex/Relex. htm* 。Lemmenmeier-Batini´c, D., N. Ljubeˇsi´c 和 T. Samardˇzi´c. 2021. 塞尔维亚语称呼形式语料库 1.0。斯洛文尼亚语语言资源库 CLARIN.SI。Ljaji´c, A. 和 U. Marovac. 2019. 通过处理塞尔维亚语中的否定规则，改进 Twitter 数据的情感分析。计算机科学与信息系统 *16*(1): 289–311。Ljaji´c, A., U. Marovac 和 M. Stankovi´c. 2019. 不同规范化方法对塞尔维亚语推文情感分析影响的比较。*Facta Universitatis*，数学与信息学系列: 683–696。Ljaji´c, A., N. Prodanovi´c, D. Medvecki, B. Baˇsaragin, J. Mitrovi´c 等. 2022. 揭示塞尔维亚民众对新冠疫苗犹豫背后的原因：基于情感的主题建模。医学互联网研究杂志 *24*(11): e42261。Ljubeˇsi´c, N., D. Boras 和 O. Kubelka. 2007. 克罗地亚语信息检索：构建简单高效的基于规则的词干提取器。Ljubeˇsi´c, N. 和 K. Dobrovoljc 2019. 神经方法带来了什么? 分析斯洛文尼亚语、克罗地亚语和塞尔维亚语形态句法标注及词元化方面的改进。载于第七届波罗的—斯拉夫自然语言处理研讨会论文集，第 29–34 页。Ljubeˇsi´c, N., K. Dobrovoljc 和 D. Fiˇser. 2015. 从已解析语料库中提取的克罗地亚语、斯洛文尼亚语和塞尔维亚语多词表达词库 Mwelex。*Informatica 39*(3)。Ljubeˇsi´c, N. 和 T. Erjavec. 2018. 词嵌入 clarin. si-embed. hr 1.0，斯洛文尼亚语语言资源库 clarin. si（2018）。Ljubeˇsi´c, N., T. Erjavec 和 D. Fiˇser 2016. 基于语料库的南斯拉夫语变音符号恢复。载于第十届国际语言资源与评估会议论文集（*LREC'16*），第 3612–3616 页。Ljubeˇsi´c, N., T. Erjavec, M. Miliˇcevi´c 和 T. Samardˇzi´c. 2017. 塞尔维亚语 Twitter 训练语料库 reldi-normtagner-sr 2.0。Ljubeˇsi´c, N., M. Espl`a-Gomis, S. Ortiz Rojas, F. Klubiˇcka 和 A. Toral. 2016. 塞尔维亚语—英语平行语料库 srenWaC 1.0。斯洛文尼亚语语言资源库 CLARIN.SI。

36 塞尔维亚语自然语言处理资源与方法综述Ljubeˇsi´c, N. and F. Klubiˇcka 2014年4月。bs、hr、srWaC——波斯尼亚语、克罗地亚语和塞尔维亚语网络语料库。载于第九届网络语料库研讨会论文集（*WaC-9*），瑞典哥德堡，第29–35页。计算语言学协会。Ljubeˇsi´c, N. and F. Klubiˇcka. 2016. 塞尔维亚语网络语料库 srwac。卢布尔雅那：*Joˇzef Stefan* 研究所。Ljubeˇsi´c, N. and F. Klubiˇcka. 2016. 塞尔维亚语网络语料库 srWaC 1.1。斯洛文尼亚语语言资源库 CLARIN.SI。Ljubeˇsi´c, N., F. Klubiˇcka, ˇZ. Agi´c, and I.P. Jazbec 2016年。用于改进克罗地亚语和塞尔维亚语形态句法标注的新型屈折词典和训练语料库。载于第十届国际语言资源与评测会议论文集（*LREC'16*），第4264–4270页。Ljubeˇsi´c, N. and D. Lauc. 2021. Berti´c——适用于波斯尼亚语、克罗地亚语、黑山语和塞尔维亚语的 Transformer 语言模型。*arXiv* 预印本 *arXiv:2104.09243* 。Ljubeˇsi´c, N., F. Markoski, E. Markoska, and T. Erjavec. 2021. 南斯拉夫语维基百科的可比语料库 CLASSLA-wikipedia 1.0。斯洛文尼亚语语言资源库 CLARIN.SI。Ljubeˇsi´c, N. and P. Rupnik. 2022. 用于区分波斯尼亚语、克罗地亚语、黑山语和塞尔维亚语的 Twitter 用户数据集——Twitter-HBS 1.0。斯洛文尼亚语语言资源库 CLARIN.SI。Ljubeˇsi´c, N., M. Starovi´c, T. Kuzman, and T. Samardˇzi´c. 2022. 塞尔维亚语合理备选项选择数据集 COPA-SR。斯洛文尼亚语语言资源库 CLARIN.SI。Ljubeˇsi´c, N., M. Stupar, T. Juri´c, and ˇZ. Agi´c. 2013年。结合现有数据集，构建克罗地亚语和斯洛文尼亚语命名实体识别模型。斯洛文尼亚语2(1)：35–57。Magner, T. 2001. 克罗地亚人和塞尔维亚人领土上的双重文字体系。《国际社会学与语言学杂志》：11–26。https://doi.org/doi:10.1515/ijsl.2001.028。Marovac, U., A. Avdi´c, and A. Ljaji´c. 2021. 塞尔维亚语停用词词典的构建。诺维帕扎尔州立大学科学出版物，*A*辑：应用数学、信息学与力学 *13*(1)：17–25。Marovac, U., A. Ljaji´c, E. Kajan, and A. Avdi´c. 2013. 塞尔维亚语文本数据的相似性搜索。*ICEST* 会议论文集：607–610。

塞尔维亚语自然语言处理资源与方法综述37 Marovac, U., A. Pljaskovi´c,A. Crniˇsanin, and E.Kajan 2012. 塞尔维亚语文档的 N-gram 分析。载于*2012*年第*20*届电信论坛（*TELFOR*），第1385–1388页。IEEE。Matthews, W.K.1952.西里尔字母字符的拉丁字母化。斯拉夫与东欧评论*30*(75)：531–548。Miletic, A. 2017. 使用维基词典构建塞尔维亚语形态句法词典。载于第六届 *´Journ´ées d'´etudetoulousaines*：语言科学中的接口，第 30–34 页。Miliˇcevi´c, M. and N.Ljubeˇsi´c. 2016.Tviterasi、tviteraˇsi还是twitteraˇsi? 生成并分析克罗地亚语和塞尔维亚语的规范化数据集。*Slovenˇsˇcina 2.0*：实证、应用与跨学科研究 *4*(2)：156–188。Miller,G.A. and C. Fellbaum.2007. 词网今昔。语言资源与评估 41：209–214。Miloˇsevi´c, N. 2012a.塞尔维亚语句子的机器情感分析。硕士学位论文 。Miloˇsevi´c, N. 2012b.塞尔维亚语词干提取器。*arXiv* 预印本 *arXiv:1209.4471* 。Milosevi´c,N.andG.Nenadi´c. 2016. 凉爽如黄瓜：构建塞尔维亚语当代明喻语料库。*arXiv* 预印本 *arXiv:1605.06319* 。Milosevi´c, N. and G.Nenadi´c. 2018. 使用自然语言处理创建塞尔维亚语当代明喻语料库。*arXiv* 预印本 *arXiv:1811.10422*。Mitrovic,J. 2014. 塞尔维亚语多词单元检测与研究的电子工具和资源。载于 *IC1207 COST*行动 *PARSEME* 第*2*届全体会议，希腊雅典，第 10–11 页。Mitrovi´c, J., M. Mladenovi´c,and C. Krstev 2015. 通过众包向塞尔维亚语词汇资源添加多词表达式。载于 在第*5*届*PARSEME* 全体会议上展示的海报。*Iasi*，罗马尼亚，第23–24 页。Mitrovi´c, J.,C. O'Reilly, M. Mladenovi´c,andS.Handschuh. 2017.用于论证采矿的修辞格本体表示。论证与计算*8*(3)：267–287。Mladenovi´c,M. 2016. 基于语言资源的情感分析信息学模型 ´。博士学位论文。

38塞尔维亚语自然语言处理资源与方法综述Mladenovi´c，M.和J. Mitrovi´c 2013年。塞尔维亚语修辞格本体。在 国际文本、语音与对话会议，第386–393页。Springer。Mladenovi´c，M.、J.Mitrovi´c 和 C. Krstev 2014年。词网的开发与维护：流程和工具。在 第七届全球词网会议论文集，第55–62页。Mladenovi´c，M.、J. Mitrovi´c、C. Krstev 和D. Vitas。2016年。面向形态丰富语言的混合情感分析框架。智能信息系统杂志*46*(3)：599–620。Mladenovi´c，M.、S.V. Stankovi´c 和 V. Paji´c。2020年。使用balkanet自动生成应用本体的两种方法。语义网与信息系统国际期刊（*IJSWIS*）*16*(2)：18–41。Mochtak，M.、P.Rupnik和N.Ljubeˇsiˇc。2022。来自波斯尼亚和黑塞哥维那、克罗地亚和塞尔维亚的情感标注议会辩论数据集parlasent-bcs。*arXiv*预印本 *arXiv:2206.00929* 。Mozetiˇc，I.和 M. Grˇcar。Smailovi c j.2016年。多语言*Twitter*情感分类：人工标注员的作用。*PLOS ONE11*(5)：e0155036。Mozetiˇc，I.、M.Grˇcar 和J. Smailovi´c。2016年。15种欧洲语言的Twitter情感。斯洛文尼亚语语言资源库CLARIN.SI。Nenadi´c，G. 2004年。创建数字语言资源。国家数字化中心综述(5)：191–30。Nikoli´c，N.、O. Grljevi´c 和 A.Kovaˇcevi´c。2020年。高等教育领域评论的基于方面的情感分析。电子图书馆*38*(1)：44–64。Nikoli´c，V.2016年。塞尔维亚共和国电子政务中非结构化数据和文档的建模与检索。博士学位论文。Odebrecht，C.、L. Burnard 和 C. Sch¨och。2021年。欧洲文学文本集（ELTeC）：2021年4月发布的包含至少50部小说的14个合集。Zenodo。Ostrogonac，S.2018。塞尔维亚语模型及其在语音和语言技术中的应用。博士学位论文，诺维萨德大学（塞尔维亚）。Ostrogonac，S.、B.Rastovi´c 和 E. Liliom。2020年。用于塞尔维亚语文本处理的Python软件包——nlpheart。科学技术评论*70*(3)：41–45。

9 Ouyang, L., J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray 等。2022。训练语言模型，使其通过人类反馈遵循指令。*arXiv* 预印本 *arXiv:2203.02155* 。Paumier, S., F. Malchok, C. Marschner, C. Martineau, C. Martınez, D. Maurel, S. Nagel, A. Neme, M. Petit, J. Stiehler 等。2002年。Unitex 3.2。Pavlovic-Lazetic, G. 和 J. Graovac，2010年。由本体驱动的概念文档分类。载于*KDIR*，第383–386页。Pavlovi´c-La˘zeti´c, G., D. Vitas 和 C. Krstev, 2004年。迈向完整的词汇识别。载于文本、语音与对话：第七届国际会议，*TSD 2004*，捷克共和国布尔诺，*2004*年*9*月*8*日至*11*日。论文集 *7*，第179–186页。Springer。Petrovi´c, D.，2020年。基于法律文档之间通过特定语言表达建立的关联对文档集合结构进行分析。博士学位论文，尼什大学电子工程学院。Petrovi´c, D. 和 S. Jani˘cijevi´c，2019年。用于训练神经网络的领域专用词嵌入矩阵。载于*2019*年人工智能国际会议：应用与创新（*IC-AIAI*），第71–714页。IEEE。Petrovi´c, D. 和 M. Stankovi´c。2019年。文本预处理方法和工具对计算文本相似度的影响。塞尔维亚数学与信息学 *34*(5)：973–994。Plisson, J., N. Lavra˘c, D. Mladeni´c 和 T. Erjavec。2008年。用于自动词形还原的涟漪式规则学习。人工智能通信 *21*(1)：15–26。Popovi´c, M. 和 M. Ar˘can。2016年。后编辑并标注错误的机器翻译语料库 {PErr} 1.0。斯洛文尼亚语资源库 CLARIN.SI。Popovi´c, Z.，2008年。塞尔维亚语文本标注程序的评估。博士学位论文。Popovi´c, Z.，2010年。应用于塞尔维亚语文本的标注器。*INFOtheca*—信息学与图书馆学期刊 *11*(2)。Puflovi´c, D., G. Velinov, T. Stankovi´c, D. Jankovi´c 和 L. Stoimenov。2016年。一种用于从医疗记录中提取信息的监督式命名实体识别方法。

40 塞尔维亚语自然语言处理资源与方法综述Reshamwala, A., D. Mishra, and P. Pawar. 2013年. 自然语言处理综述. *IRACST* 工程科学与技术：国际期刊（*ESTIJ*）*3*(1): 113–116 . Samardˇzi´c, T., N. Ljubeˇsi´c, and M. Miliˇcevi´c. 2015. 区域语言数据倡议（reldi）. Samardˇzi´c, T., M. Starovi´c, ˇZ. Agi´c, and N. Ljubeˇsi´c. 2017. 塞尔维亚语依存关系，与克罗地亚语和其他斯拉夫语进行比较. ˇSandrih, B. 2019年. 基于词汇特征、表情符号和非正式缩写的短信情感分类. 塞迪卡计算学期刊 *13*(1-2): 081p–096p . ˇSandrih, B., C. Krstev, and R. Stankovi´c 2019年. 塞尔维亚语三个命名实体识别系统的开发与评估：以人名为例. 载于自然语言处理最新进展国际会议论文集（*RANLP 2019*），第 1060–1068 页。Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019年. DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量. *arXiv* 预印本 *arXiv:1910.01108* . Seˇcujski, M. and V. Deli´c. 2008年. 用于塞尔维亚语自动词性标注的软件工具. 应用语言学 *1*(9): 97–103 . Seˇcujski, M.S. and A.D. Kupusinac. 使用 HMM 对塞尔维亚语文本进行自动形态标注. Sekine, S. 2004. 命名实体：历史与未来. 纽约大学项目说明: 4 . Stankovi´c, R., M. Koˇsprdi´c, M.I. Neˇsi´c, and T. Radovi´c 2022. 塞尔维亚语古典小说的情感分析. 载于第二届情感分析与语言链接数据研讨会论文集，第 31–38 页。Stankovi´c, R., C. Krstev, I. Obradovi´c, A. Trtovac, and M. Utvi´c 2012年. 用于增强多语种电子期刊数字图书馆检索的工具. 载于第八届语言资源与评估国际会议论文集（*LREC 2012*），*2012*年*5*月，土耳其伊斯坦布尔，第 1710–1717 页。Stankovi´c, R., J. Mitrovi´c, D. Joki´c, and C. Krstev 2020年. 用于塞尔维亚语辱骂性言论检测的多词表达. 载于多词表达与电子词典联合研讨会论文集，第 74–84 页。

1 Stankovi´c、R.、I. Obradovi´c、C. Krstev 和 D. Vitas 2011。使用多用途工具制作多词单位的形态词典。载于计算语言学应用会议论文集，*2011*年*10*月，波兰雅赫兰卡。Stankovi´c、R.、B. Trivi´c、O. Kitanovi´c、B. Blagojevi´c 和 V. Nikoli´c。2011年。Geolissterm 术语词典的开发。*INFOtheca*-信息学与图书馆学期刊 *12*(1)。Stankovi´c、R.、C. Krstev、B.ˇS. Todorovi´c 和 M. ˇSkoric。2021年。塞尔维亚语 ELTeC 文集的标注。*Infotheca*-数字人文学期刊 *21*(2)：43–59。Stankovi´c、R.、B. ˇSandrih、C. Krstev、M. Utvi´c 和 M. Skoric，2020年5月。基于机器学习和深度神经网络的塞尔维亚语词形还原与形态句法标注。载于第十二届语言资源与评测会议论文集，法国马赛，第3954–3962页。欧洲语言资源协会。Tanasijevi´c，I. 2019年。迈向文化遗产文档的自动标注。*IPSI* 高级研究汇刊，*TAR 15*(1)。Thoppilan，R.、D. De Freitas、J. Hall、N. Shazeer、A. Kulshreshtha、H.T. Cheng、A. Jin、T. Bos、L. Baker、Y. Du 等。2022年。Lamda：用于对话应用的语言模型。*arXiv* 预印本 *arXiv:2201.08239* 。Tiedemann，J. 2012年。OPUS 中的平行数据、工具与接口。载于$Lrec$，第2012卷，第2214–2218页。ˇTodorovi´c、B.S.、R. Stankovi´c、C. Krstev 和 M.I. Neˇsi´c，2021年。塞尔维亚语命名实体识别及其他：古语与现代语的交织。载于深度学习自然语言处理方法与应用——自然语言处理最新进展国际会议论文集（*RANLP 2021*），第1252–1260页。Tomaˇsevi´c，A.D. 2018年。矿业项目文档管理模型的开发。博士学位论文。Ulˇcar，M. 和 M. Robnik-ˇSikonja，2020年。Finest BERT 和 CroSloEngual BERT：多语言模型中的少即是多。载于文本、语音与对话：第*23*届国际会议，*TSD 2020*，捷克共和国布尔诺，*2020*年*9*月*8*日至*11*日，论文集 *23*，第104–111页。Springer。Utvi´c，M. 2011年。当代塞尔维亚语语料库的标注。载于 *INFOtheca '12* 会议论文集，第36–47页。

ate 42 塞尔维亚语自然语言处理资源与方法综述Utvi´c, M. 2014. 现代塞尔维亚语参考语料库的构建。博士学位论文。Utvi´c, M., R. Stankovi´c, A. Toma˘sevi´c, M. ˘Skori´c, and B. Lazi´c. 2019. 基于外部词汇资源使用的语料库检索：通过网络服务实现。*Vukovo vreme* 第*48/3*卷：斯拉夫学者科学会议——塞尔维亚语及其资源。Vasiljevi´c, N. 2015. 塞尔维亚语法律文本的自动处理。博士学位论文。Vitas, D., S. Koeva, C. Krstev, and I. Obradovi´c 2008. 穿越词典的环球之旅。载于第*27*届词汇与语法国际研讨会论文集，第249–256页。Vitas, D. and C. Krstev. 2006. 文学与对齐文本。多语言读本：148–155。Vitas, D. and C. Krstev. 2009. 塞尔维亚语与sstbi。*SNTPI '09*——科学、技术与商业信息系统科学专业会议。´Vitas, D., C. Krstev, and E. Laporte. 2006. 双语文本的准备与利用。*Lux Coreana*（1）：110–132。Vitas, D., C. Krstev, I. Obradovi´c, L. Popovi´c, and G. Pavlovi´c-La˘zeti´c 2003a. 塞尔维亚语书面文本处理资源与基本工具概述。载于巴尔干语言资源研讨会论文集，第一届巴尔干信息学会议。Citeseer。Vitas, D., C. Krstev, I. Obradovi´c, L. Popovi´c, and G. Pavlovi´c-La˘zeti´c 2003b. 塞尔维亚语书面文本处理：资源与基本工具概述。载于巴尔干语言资源与工具研讨会，第21卷，第97–104页。Vitas, D., P. Ljubomir, K. Cvetana, O. Ivan, P.L. Gordana, and S. Mladen. 2012. 数字时代的塞尔维亚语。*META-NET* 白皮书系列，*G. Rehm*、*H. Uszkoreit*（编）。Vitas, D., G. Nenadi´c, and C. Krstev 1998. 与17种语言对齐的柏拉图《理想国》塞尔维亚语译本电子版。载于东遇西——多语言资源概要，TELRI Association e.V.，Institut fur deutsche Sprache，曼海姆。Vitas, D. and G. Pavlovi´c-La˘zeti´c. 2008. 塞尔维亚语命名实体识别的资源与方法。*INFOtheca*——信息学与图书馆学期刊9。

Vujani´c，M. 编，2007 年。塞尔维亚语词典。Matica srpska。

Vujiˇci´c-Stankovi´c, S. 2016年. 本体驱动的信息抽取：塞尔维亚语模型。博士学位论文。

Vujiˇci´c-Stankovi´c，S.、C.Krstev和D. Vitas 2014年。使用烹饪领域的术语丰富塞尔维亚词网和电子词典。载于第七届全球词网会议论文集，第 127–132 页。

Vukovi´c, T. 2020年。《托尔拉克方言口语语料库 1.0》（转录）。斯洛文尼亚语语言资源库 CLARIN.SI。

Wankhade,M.、A.C.S. Rao 和C.Kulkarni。2022。《情感分析方法、应用与挑战综述》。人工智能评论：1–50。