俄语网络仇恨言论语料库研究与构建

温昕
西安电子科技大学外国语学院
中国，西安 710071
wenxin@xidian.edu.cn

郑敏娇
国防科技大学信息通信学院
中国，武汉 430010
mjzzheng@126.com

摘要

近年来，网络科技的飞速发展在为整个社会带来极大便利的同时，也加剧了仇恨言论的传播。仇恨言论可能会构成网络暴力，诱发仇恨性的犯罪行为，对社会公共文明和网络空间秩序造成极大的威胁。因此，对网络仇恨言论进行主动的监管和制约具有重大意义。而当前学术界针对俄语的网络仇恨言论研究不足，尤其缺乏俄语网络仇恨言论语料库，这极大地限制了相关技术和应用的发展。2022年俄乌冲突爆发以后，对于俄语网络仇恨言论语料库的研究与构建显得更加迫切。在本文中，作者提出了一种细粒度的俄语网络仇恨言论语料库构建及标注方案，并基于该方案首次创建了包含20476条文本数据，具有针对性、话题统一的俄语仇恨性言论语料库。

关键词： 俄语 ； 语料库 ； 仇恨言论

An Russian Internet Corpus for Hate Speech Detection

Xin Wen
School of Foreign Languages,
Xidian University,
Xi'an, China 700071
wenxin@xidian.edu.cn

Minjiao Zheng
School of Information and Communication,
National University of Defense Technology,
Wuhan, China 430010
mjzzheng@126.com

Abstract

With the rapid development of network society, the spread of hate speech has become increasingly serious, which attracts more and more people's attention. Hateful speech may induce cyber violence, hate crimes, and cause great harm to society. Therefore, it is necessary to supervise online hate speech, and thereby reducing the potential harm. However, there is a lack of relevant research on Russian online hate speech, especially the research on Russian hate speech corpus, which greatly limits the development of the relevant researches. To solve this problem, this paper creates a Russian online hate speech corpus that has a unified topic, and achieves multi-dimensional labeling on the corpus.

Keywords: Russian , Corpus , Hate Speech

1 引言

第二十一届中国计算语言学大会论文集，第694页-第703页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

信息爆炸、数据井喷是21世纪社会现状的显著特征。在高速发展的信息流时代，社交网络开始全方位地影响我们的日常生活。全球约有39.6亿人使用社交媒体，占全球77.7亿总人口的50.64%；另有数据表明，有83.36%的互联网用户都在使用社交媒体，网络媒体已逐渐成为人们交流观点、表达情感的最重要途径。看似自由的网络空间实则具有虚拟性、欺骗性、随意性、复杂性，难以监管，网络社交媒体常常成为仇恨性言论的爆发地。

所谓仇恨言论（Hate Speech），是指在仇恨心态的引导下，基于民族、种族、国籍、宗教、性别、年龄和身体状况等身份特征，公开发表，来表达对他人或某些群体的仇恨、欺凌、贬低、侮辱、歧视、威胁或煽动暴力的言论(Erjavec and Kovačič, 2012)，其往往针对于弱势群体进行攻击，涉及到许多社会存在的焦点问题，如种族歧视、性别对立、宗教信仰、性取向等。例如，2014年的乌克兰反政府示威运动（Евромайдан），以及2022年的俄乌冲突爆发以来，俄语网络社交媒体中就产生了大量的仇恨言论。

虽然对于仇恨言论的定义和定义方式各不相同，但多数专家和学者认为，仇恨言论具有极强的负面影响力，不属于言论自由所应当保护的言论，必须受到限制(Василенко, 2019)。仇恨性言论在网络上的传播可能会引发恶劣的结果，诱导人们（尤其是青少年）实施仇恨性犯罪，影响社会安定，危害公共安全，甚至间接导致国际冲突升级。为了制约网络仇恨性言论的产生与传播，各大社交网络平台也制定了相应的规则。例如Twitter、微博等平台都设有专门的内容审查员手动检测和审核包含此类语言现象的评论或帖子。为了提升检测效率，研究自动化、低成本的网络仇恨言论检测方法变得十分必要。而语料库作为语言学研究的基础资源，是研究仇恨言论检测方法的必要条件。当前学术界针对俄语的网络仇恨言论缺乏相关的研究，尤其缺乏俄语网络仇恨言论语料库。因此，研究并建立具有针对性的、话题统一的俄语网络仇恨言论语料库具有重要的理论和实践意义。

针对此问题，本文以乌克兰民族和政治问题为主题，构建了一个俄语仇恨言论语料库，并探讨了语料库的标注方法。本文选取了YouTube这一社交网络平台中的话题：“乌克兰反政府示威运动（Евромайдан）”，对相关视频下方的评论文本进行抓取和收集，以此构建语料库。同时，我们提出了多维度、细粒度的标注方法，从是否含有仇恨言论，有无攻击性、冒犯性、讽刺性、刻板印象、辱骂性等维度及仇恨的强弱程度对评论内容进行标注。此种多维度的标注方法能够使文本内容中的细微差别得到更全面的展现，为仇恨言论的识别和界定提供更精细的参考标准，能够较好地平衡标注的正确性和主观性。最后，对标注结果和方法进行分析与讨论。我们发现，在本语料库所选取的乌克兰民族与政治的主题中，仇恨言论在攻击性和冒犯性两个维度上表现得最为明显，但依然需要考虑刻板印象和讽刺性等隐式煽动的仇恨表达方式。

我们创建的俄语网络仇恨言论语料库包含20476条评论数据，该语料库可作为俄语网络仇恨言论检测技术的测试集和训练集，是相关研究的基础。

2 相关工作

当前在俄语研究范畴内，对于仇恨言论的研究主要是将其作为一种语言现象进行研究，探索仇恨言论在传播理论、多语言环境等语用方面的问题(Василенко, 2019; Хроменков, 2016; Шарнин et al., 2018)，但并未给出关于仇恨言论的定性或定量的描述方法，也没能形成俄语网络仇恨言论相关的权威语料库。在计算语言学和人工智能飞速发展的背景下，语料库缺失这一基础问题极大地限制了俄语相关研究的发展。

目前，对于仇恨言论的语料库研究大多基于英语。Kennedy等(2018)组织语言专家根据编码类型学标记社交网络中的帖子，形成了仇恨语料库GHC（Gab Hate Corpus）；Assimakopoulos等(2020)提出了针对马耳他地中海移民危机、性小众群体（LGBTIQ+）等问题的多层语料库标注方案；Klubička等(2018)强调了现有的仇恨言论检测方式存在的问题，并在语料库的基础上对仇恨言论检测之外的问题进行探索；Huang等(2020)在语料库中标注了文本发布人的年龄、国家、性别和种族/民族等因素，用以分析言论偏差和评论的人口可预测性。

除英语以外，阿拉伯语、意大利语等小语种中也有部分涉及仇恨言论语料库的研究。Alakrot等(2018)基于YouTube数据构建了一个阿拉伯语评论语料库，并将其用于辱骂性语言的检测；Alhuzali与Abdul-Mageed(2018)在Twitter上获取了阿拉伯语的评论内容，构建了语料库和仇恨词典；在意大利语中，针对仇恨言论语料库和仇恨言论检测方法相关的研究有Del Vigna等(2017)；Maisto等(2017)；Bosco等(2018)。

第二十一届中国计算语言学大会论文集，第694页-第703页，南昌，中国，2022年10月14日至16日。

从语料库标注的角度来看，Kwok和Wang(2013)的研究对Twitter文本的攻击性进行了分类，等级为1到5；Ross等(2017)为语料库的标注设计了两个标签，即仇恨言论（是/否）和进攻性，等级从1到6；Del Vigna等(2017)的研究使用了包括无仇恨/弱仇恨/强仇恨三个等级的标签。以上工作表明，简单的二元标签无法满足分析仇恨言论这一复杂的需求，并且可能会影响后续研究和进一步分析的准确性。因此，本文同样采用了非二元标签语料库标注方法，并在此基础上引入了多个注释类别，建立了多维度的语料库标注方案。

从语料库收集的角度看，现有相关研究大多通过仇恨类别或一组典型的仇恨词语来收集语料数据(Del Vigna12 et al., 2017; Sanguinetti et al., 2018)，而Waseem和Hovy(2016)在收集语料时将仇恨词语和经常与仇恨言论一同出现的中性词语结合在一起，可以识别更宽泛的仇恨言论的表达方式。也就是说，通过典型的仇恨词来收集语料库的方法存在片面性的问题。因此，为了解决该问题，本文采用了基于主题的语料数据收集方法，以仇恨言论的高发话题——乌克兰民族和政治问题中的相关网络评论作为语料来源。

总而言之，与上述相关工作相比，本文致力于提出一种更新颖、更细粒度的语料库构建及标注方案。该方案能够用以表示仇恨言论现象的多个方面，因此具有更强的挑战性。

3 语料库的创建

3.1 为什么选择民族政治问题作为语料库的语料来源

由于民族和种族问题是因历史、政治等复杂因素造成的问题，由来已久并且十分尖锐，容易引起言语冲突和言语对抗。敏感的政治问题和暴力冲突事件更容易引发民族和不同派别之间的矛盾。在网络社会中，有关此类事件的评价与讨论大多都是负面的，因此我们在收集语料时将针对乌克兰民族和政治问题的网络评论作为本语料库的语料来源。

我们选取了2013年至2014年发生在乌克兰的“反政府示威运动（Евромайдан）”这一事件为主题，是因为此次运动造成了极大的影响。据统计，为期93天示威造成至少125人死亡，1890多人受伤，65人失踪。由该运动引起的话题和相关视频在网上掀起了轩然大波，引起亲欧派和亲俄派、乌克兰族和俄罗斯族的激烈争论。因此，我们收集了YouTube社交网络平台中与此次运动相关的视频下方的评论文本。

3.2 为什么提出多维度的标注方案

我们将上文提到的评论文本进行整理，获得了20476条评论数据，从中随机选择了1220条评论作为标注子集。我们首先对标注子集进行机器翻译和人工校对。带有翻译内容的数据集分别由语言学专家和以俄语为母语的标注志愿者对其进行标注。标注的具体方案在研究中都将进行详细介绍。我们将对标注的结果进行比照分析，思考标注原则的制定是否合理，分析出现标注偏差的原因，提出日后改进的方向。

我们的标注内容包括是否含有仇恨言论、有无攻击性、冒犯性、讽刺性、刻板印象、辱骂性等维度和仇恨的强弱程度。这样的设计是为了使专家、标注者和我们能够深刻理解和体会仇恨言论的定义和范畴，通过各个含义和角度的交叉验证，帮助我们对文本的仇恨内容进行更加理性的判断。另外，我们建立的语料库话题统一，在内容上具有很强的针对性，能够为仇恨言论自动检测系统的训练提供更加契合的基础支持，并为其他相关研究奠定语料库基础。

4 标注方案

仇恨言论的鉴别是一项非常具有挑战性的任务，不可避免地会受到判断者个体认知和主观思维的影响(Waseem and Hovy, 2016; Ross et al., 2017)。Weber(2009)借助欧洲人权法院的运作方式来说明鉴别仇恨言论的困难之处，并且强调，在划定言论的合法性、非法性的界限时，并没有单一的标准，而是需要借助一组综合性的考量标准，并且进行逐案分析。因此，我们尝试不仅仅鉴别言论中是否存在仇恨性，还参考其他的辅助参数来标注每一条评论以减少标注者的犹豫和不确定，使其能够对每一条数据做出更理性、更准确的判断。

我们制定了一套标注指南，尝试将所有标注项包含在一个统一的框架之中。我们设计的标注项目除了仇恨言论（HS，是/否）以外，还包括攻击性（aggressiveness，强/弱/无）、冒犯性（offensiveness，强/弱/无）、讽刺性（irony，有/无）、刻板印象（stereotype，有/无）和辱骂性（abusive，有/无）。

考虑到此注释任务需要面对标注主体（人类）本身的复杂性，以及该任务固有的复杂性，我们分别邀请了两组人员对同一数据进行了标注。我们将由专家组成的A组标注结果和由俄语母语人组成的B组标注结果进行对比，发现两组标注结果也存在一些分歧，根据这些结果，我们深入讨论了发生分歧的可能的原因。

4.1 标注项目

以下是我们对标注的项目所给出的定义和示例：

仇恨言论（HS，是/否）——对于评论是否包含仇恨言论的判定，主要从两个元素上辨别：

- **目标**——仇恨言论的目标需是一定的群体（俄罗斯人/乌克兰人/少数民族/亲俄派/亲欧派）或者是作为这些群体成员的个人。
- **言语意图**——话语的言外之意(Searle and Searle, 1969)。我们需要了解言语的隐含意图，捕捉言语中对目标的渗透、煽动、刺激以及仇恨或暴力的迹象，或是言语中使目标群体失去人性化，丧失合法性，对目标进行伤害或恐吓。

这两个元素在评论数据中同时存在对于确定评论中是否包含仇恨言论至关重要，例如：отстреливайте хохлов да поскорее да боевыми да наглухо валите их（快发动军队向乌克兰人开枪，狠狠地击垮他们）。如果在评论中不包括任何一个元素，那么就认为这条评论中不包含仇恨言论。下面我们对其余的标注类别分别作简要说明：

- **攻击性（aggressiveness，强/弱/无）**——关注用户的攻击、侵略、伤害等直接行为，甚至是煽动对既定目标实施各种形式的暴力行为。如果存在这种意图，还可以对其程度进行强/弱的划分。例如，言论中暗示歧视或将歧视合法化的态度会被视为是攻击性较弱的言论：по русски пожалуйста（请说俄语），这里指对乌克兰人说俄语规范性的一种轻视和瞧不起；而如果提及暴力行为，无论是明示还是暗示，都被认为具有强烈的攻击性：хороший мусор мертвый мусор вырезать их семьи отрезать головы и в футбол играть（好垃圾死垃圾砍了他们的家人砍掉他们的头踢足球）。
- **冒犯性（offensiveness，强/弱/无）**——与攻击性相反，冒犯性关注的是评论内容对给定目标的间接伤害和潜在的影响，对冒犯性的程度也可以进行强/弱的划分。例如，认为给定目标具有典型的人的缺陷时，则评论具有弱冒犯性：наталия вы глупы половина украинцев говорит на русском но при этом русскими не становятся（娜塔莉亚您真是愚蠢，半数乌克兰人都说俄语，但他们也不是俄罗斯人）；而如果目标面临粗暴或侮辱性的表达时，则该评论内容将被标记为具有强冒犯性：ну ты и дура как я посмотрю（好吧，在我看来你就是个傻瓜）。
- **讽刺性（irony，有/无）**——该术语作为通用术语，涵盖了讽刺、幽默、反讽等讽刺意味的细微差别（Bosco等，2013：4159）。在语料库中，对讽刺性的衡量是二元的，即有/无。在数据标注中引入讽刺性源于对数据的初步观察。我们在评论数据中看到了讽刺，尤其是反讽对于仇恨性的削弱或间接表达仇恨性内容的现象非常普遍，尤其是俄罗斯人的语言风格和言语习惯，他们会经常引入一些幽默风趣的表达，所以我们引入讽刺性作为一个仇恨性内容判断的参考项。例如：федорову нужно писать сценарии для голливуда（费多罗夫需要为好莱坞写剧本）；по мне так всё логично эти мирные демонстранты сожгли много зданий взяли в заложники людей захватывают административные здания и не идут на мир думаю многие заслужили пулю от снайпера в лоб（对我来说，一切都如此合乎逻辑，这些和平示威者烧毁了许多建筑物，劫持了许多人质，攻占了行政大楼，就是不要和平，我觉得他们中的很多人都该被枪毙）。
- **刻板印象（stereotype，有/无）**——判断评论内容中是否含有对特定目标明显的或隐含的一种概括或固定的看法，即认为整个目标群体都具有该特征，从而忽视个体差异。在本研究中，刻板印象主要表现为对目标群体的偏见。我们通过对数据的观察发现，对少数群体的仇恨也常常以偏见为特征。在我们的语料库中，对刻板印象的衡量也是以二元的有/无

作为判断标准的。例如：прикол в том что украине не предлагают шенген там своих нахлебников хватает（搞笑的是乌克兰不给申根签证，因为那里的寄生虫已经够多了）。

- **辱骂性（abusive，有/无）**——主要指语言中的脏话，禁忌语。由于网络环境相对自由，所以在网络言论中辱骂性非常普遍。仇恨言论在多数情况下表现为辱骂性言论，例如：пшел от сюда дичь（滚吧，畜生）。

4.2 标注程度

在对标注后的数据结果进行更详细的观察时，我们发现被标注为仇恨言论的数据在强度和危害程度上都有着较大的差异，所以仅仅区分评论内容中是否含有仇恨言论是不能精确地反映其仇恨性的。因此，我们引入了对仇恨程度的考量，即用"煽动强度"这个概念来解释不同类型的仇恨言论，甚至是言语中的暴力行为。我们发现，这种程度上的界定对于我们正确理解和深度挖掘评论的含义有着非常大的帮助。我们将仇恨言论的煽动性定义为5个等级，其值分别为0-4级，如果不含有仇恨言论，则值为0。

- **0 级**：完全没有煽动性，文本的内容有些矛盾，尽管评论文本可能被注释为具有攻击性、冒犯性或其他形式，但并不包含仇恨言论：насилие это классно вы че сука（暴力很酷你这个婊子）。

- **1 级**：没有明确的煽动性，但说话人将某些不良特征或品质归因于目标群体。有时，他们认为这些负面特征可能会对说话人和读者构成威胁，这些评论更类似于基于刻板印象的侮辱或判断：они только скакать жрать срать спать могут и им этого вполне хватает для счастья （他们只会跳，吃，拉，睡，有了这些他们就足够幸福了）。

- **2 级**：没有明确的煽动，但是文本中表达的行为旨在使目标群体失去人性化，丧失合法性，或者声称给予他们的基本权利是不公正的特权，亦或是声称目标群体的这些权利会伤害说话人和读者，因此不应再被授予。这些言语行为不是呼吁暴力，但会引起对目标群体的厌恶或仇恨：какая мразь не дай бог придут к власти в россии опять реки крови потеря курил мы и так теряем наших братьев на украине каждый день каждый час заявляю официально хакамада враг снг кто ее поддерживает враг эта гадость не умеющая одеваца откровенная дура и демогог（什么渣滓，上帝保佑，千万别再让俄罗斯掌权了，我们经历着血流成河，在乌克兰我们每天、每个小时都在失去我们的兄弟，我正式宣布哈卡马达（Ирина Муцуовна Хакамада）是独联体的敌人，支持她的人是彻头彻尾的傻瓜、煽动者）。

- **3 级**：明确煽动暴力或歧视性行为，但说话人拒绝承担这些行为的责任，只为这些行为辩护或表达希望发生这种行为的意愿：понаражали долбаёбов майдановцев срочняков которые свою страну защищают просто убивают твари вы позорные надеюсь сдохните скоро（Maidan示威者，这些蠢兵，他们靠杀人保卫自己的国家，可耻畜生，我希望你们快点死）。

- **4 级**：明显煽动暴力或歧视行为；评论者公开提议或呼吁采取这些行动，并宣称自己已准备好执行这些行为，或参与实现这些行动：отстреливайте хохлов да поскорее да боевыми да наглухо валите их（快发动军队向乌克兰人开枪，狠狠地击垮他们）；я конечно не провокатор но я бы на месте беркута стрелял боевыми дабы пресечь этот беспредел（我不是挑衅者，但为了阻止这场混乱，我如果是金雕，我就会开枪）。

综上所述，完整的标注方案由以下类别和标签组成：

1. **hate speech（仇恨言论）：no – yes（无-有）**
2. **aggressiveness（攻击性）：no - weak – strong（无-弱-强）**
3. **offensiveness（冒犯性）:no - weak - strong（无-弱-强）**
4. **irony（讽刺性）：no - yes（无-有）**
5. **stereotype（刻板印象）：no - yes（无-有）**
6. **intensity（程度）：0 - 1 - 2 - 3 – 4**

第二十一届中国计算语言学大会论文集，第694页-第703页，南昌，中国，2022年10月14日至16日。

评论内容	仇恨言论	攻击性	冒犯性	讽刺性	刻板印象	辱骂性	仇恨程度
(1) а нах мне твоя россия россия всегда претендует на землю украины и в добавок уничтожает украинцев так было всегда так что хватит украину приписывать к россии у украины свой путь и нелезте к нам 我才不需要你们俄罗斯呢，俄罗斯觊觎乌克兰的土地，此外还杀乌克兰人，一直如此，不要再往妄想把乌克兰划入俄罗斯了，乌克兰有自己的发展道路，不要把手再伸向我们	有	无	弱	无	有	无	2
(2) пропаганда украинских наци 乌克兰纳粹的宣传	有	无	强	无	无	无	1
(3) молодцы украинцы приятно видеть то как они боролись за свои права и свободы и били поганых мусоров слава украине 干得好，乌克兰人很高兴看到他们如何为自己的权利和自由而战，打击糟糕的垃圾，光荣属于乌克兰	有	强	强	无	无	有	3
(4) давно пора эту мразоту майдановскую боевыми отстреливать 早该用武力打击这个Maidan败类了	有	弱	强	无	无	有	4
(5) менты не люди люди не менты 警察不是人，人们不是警察	有	无	强	有	无	有	1
(6) хочу что бы россия поступила с украиной как с грузией закрыть границу намертво и посмотреть как евродемократия будет процветать 我希望俄罗斯像对待格鲁吉亚一样对待乌克兰，严格关闭边界，看看欧洲民主将如何繁荣	无	弱	无	有	无	无	0
(7)я бы в вас кинул уебков если ..... 如果.....我会向你们这些混蛋扔	有	强	强	无	无	有	4
(8)надо было чтоб они тебе голову проломили чтоб другие такие как ты не рождались 他们应该打断你的脑袋，以免再生出像你这样人	有	强	无	无	无	无	3

Table 1: 标注示例

4.3 标注示例

表1中的示例显示了我们在语料库中如何应用4.1和4.2提及的类别和程度进行标注。如上所述，在我们的方案中，评论文本的仇恨性程度与我们标注的强度直接相关：如果文本中不存在仇恨言论，则其强度等于0，否则其强度的范围为1级到4级。除了强度这个标注项以外，所有其他的标注项都是相互独立的，我们给出的所有标注项既可以单独存在于句子中，也可以同其他标注项一起出现在句子中。因此，一条评论文本中可能包含仇恨言论，但同时几乎不包含或只含有一种其他的现象和含义（见表1中的评论2）；仇恨言论可能伴随着许多其他现象一起存在（评论1、3、4、5）；一条评论中也可能不包含仇恨言论，但是含有其他的现象（评论6）。

示例（1）表达了说话人对俄罗斯一贯行为的印象，认为俄罗斯总是掠夺土地、伤害人民，同时表达了对俄罗斯此种行为的厌恶，期望俄罗斯停止这种行为。此评论中还有对俄罗斯行为的刻板印象，同时话语中隐含俄罗斯的行为会对说话人的群体造成威胁和伤害，会引起对目标群体（俄罗斯人）的厌恶或仇恨，因此，选择将其注释为仇恨言论的强度为2级。在评论（2）中，说话人认为某种行为是乌克兰的纳粹行为，是对群体的一种言语冒犯，“纳粹”这一词反映了对目标群体侮辱性的判断，认为目标群体隐含的这种负面特征可能会对说话人及其群体构成威胁，无明显的煽动性言论，因此仇恨程度为1级。第（3）条评论中“打击糟糕的垃圾”这一言论内容中包含了攻击的行为（“打击”），但这是对句子中主语（“乌克兰人”）暴力行为的一种赞扬，而说话者本身并未表现出参与暴力的意愿。文本中同时包含了言语的冒犯（“垃圾”）和辱骂性词汇（“垃圾”），表达了对暴力的赞扬，因此标注仇恨性的等级为3。第（4）条评论中“败类”是对反政府示威人员的侮辱性的形容，具有辱骂性和冒犯性。本条评论和第（3）条评论中都有“打击”一词，但在本句中，这一行为是说话人对暴力的直接呼吁，并且强调“用武力打击”，可见呼吁暴力的程度非常高，因此仇恨性标记为4级。第（5）条评论带有明显的讽刺意味，该评论运用回文的修辞方式表达了对“警察”的强烈不满，并且辱骂他们“不是人”，具有冒犯性、辱骂性，不具有攻击性，因此仇恨性等级标注为1级。在评论（6）中，不存在仇恨性的言论，说话人“希望”俄罗斯对乌克兰实施“严格关闭边界”政策，这属于一种对目标群体实施某种政策的期许，带有较弱的攻击性，但并没有呼吁暴力或者伤害、侮辱等行为，所以不属于仇恨言论，“看看欧洲民主将如何繁荣”属于一种带有讽刺意味的表达，因此含有讽刺性。

以上描述的示例主要是为了能使读者清楚我们在语料库标注时所采用的判断标准。同时，这些标注的选择也突出了我们研究的关键点——首先要明确各个维度标注项目的定义，然后正确选择标注项，最后实现维度和程度上的精准标注。

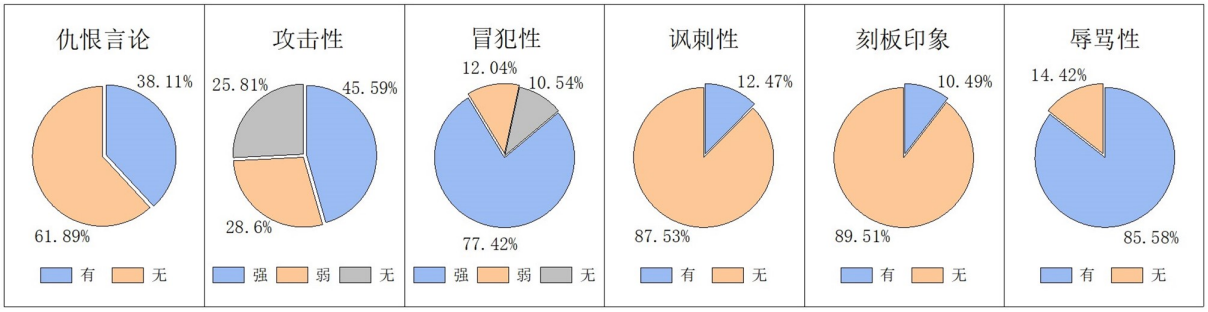


Figure 1: 各标注项在语料库中的占比

5 结果与讨论

5.1 标注结果分析

在本文中，我们对获取到的20476条评论内容中的1220条随机评论文本进行释义和标注，经过统计得到了仇恨言论在总体数据上的分布情况，以及各个维度的标注项在仇恨言论中的分布情况。考虑到我们的标注语料库是为仇恨言论自动检测研究所服务的，因此，在这里我们以仇恨言论为中心，主要研究与仇恨言论相关的其他维度的标注项在评论文本中的表现，并对这些标注项与仇恨言论的关系进行观察和分析。通过对数据的观察，我们发现在以民族与政治问题，特别是本文中选取的反政府示威运动事件为主题的语料库中，仇恨言论多以冒犯性（占评论数据的88%）、辱骂性（86%）和攻击性（74%）作为其主要的表达方式，而讽刺性（12%）和刻板印象（10%）的占比较低（如图1所示）。正如我们之前探讨过的，网络空间具有虚拟性、欺骗性、随意性和复杂性，这使得网络社交平台常常成为仇恨性言论的爆发地。人们在网络上即使是发表了极端的、情绪化的言论，明确地煽动实施暴力或歧视性的行为，似乎也不需要付出什么代价和成本。特别是我们所选取的民族和政治这类话题更容易引起极端情绪的宣泄，甚至是直接的言语对抗和辱骂。因此，在仇恨言论中，带有攻击性和冒犯性的言论占多数，成为人们表达仇恨的主要方式。

随后，我们研究了仇恨言论的强弱程度与攻击性、冒犯性、辱骂性，以及讽刺性和刻板印象之间可能存在的联系。图2中的结果显示了这些标注项在1至4级仇恨强度上的分布。

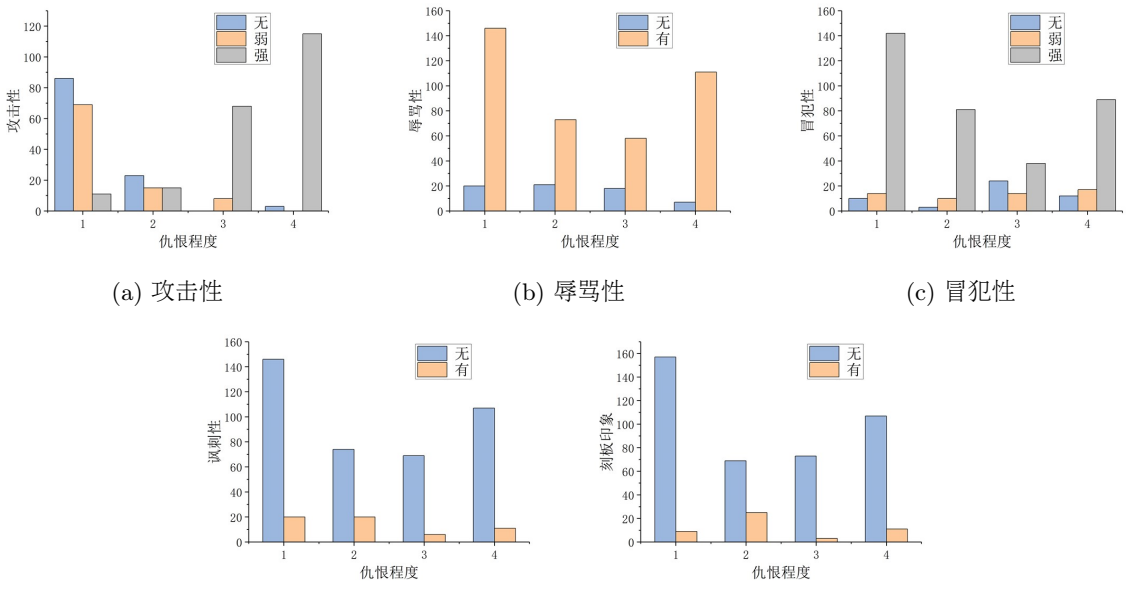
关于攻击性，在图2-a中我们可以观察到，在较低强度的仇恨等级（1级和2级）中，文本中不存在攻击性或仅存在弱攻击性的频率非常高，但强攻击性的文本出现在低等级仇恨言论中的概率就非常低了。这表明，评论文本在较高的仇恨等级（3级和4级）中，大多数都具有强攻击性。因此，在对仇恨言论及其煽动性进行定义时，需要着重考虑攻击性。整体上来看，无论哪种仇恨程度，攻击性言论的出现频率都很高，并且攻击性的强弱程度和出现的频率与仇恨性的强弱程度基本一致。这些发现与我们上述对仇恨言论的表达方式的理解完全一致。

在图2-b中，我们能够看到，与仇恨言论关系同样十分密切的辱骂性言论，其分布几乎涵盖仇恨程度的所有级别，并且所占比例非常高，不含有辱骂性的仇恨言论占比非常低，我们几乎可以得出结论，认为辱骂性和仇恨性具有极高的吻合度和一致性。

从图2-c中我们能够观察到，冒犯性在所有仇恨的等级中均有分布，冒犯性几乎跨越了所有仇恨的强弱等级。然而也有一些规律可循：冒犯性在1级和4级上所占比例更多。这是因为，在标注规则中我们规定仇恨等级为1的是没有明确的煽动性，评论者将不良特征或品质归因于目标群体。这些评论更类似于基于刻板印象的侮辱或判断，也就是对目标群体的人格进行侮辱，甚至是谩骂。但这些言论行为没有煽动对目标群体实施暴力等过激行为，因此很多冒犯性的言论都被归为仇恨程度较弱的1级。在4级仇恨中，冒犯性出现的频率也比较高，这可能是由于在4级仇恨中，出现频率最高的冒犯性和攻击性往往同时出现。

通过对比图2-b和图2-c我们发现，辱骂性和冒犯性，特别是强冒犯性在4个仇恨等级中的分布形势非常一致，这是因为绝大多数强冒犯性的言论都具有辱骂性（当然，不排除少数冒犯性言论不具有辱骂性），所以二者的分布态势基本一致。

在对图2-d和图2-e的观察中我们发现，虽然讽刺性和刻板印象在仇恨言论中的占比不高，但其二者依然对仇恨言论的表达具有很重要的影响，不容忽视。情感文本中的情感分为显式情感和隐式情感。显式情感是指在一个情感文本中包含明显的情感词语，如高兴、漂亮、很棒这



(a) 攻击性　(b) 辱骂性　(c) 冒犯性

(d) 讽刺性　(e) 刻板印象

Figure 2: 仇恨程度1-4级在各标注项中的分布

些词语。而隐式情感是指在一个情感文本中并不包含情感词语，比如："您真能说，把黑的说成白的。"带有讽刺性和刻板印象的仇恨言论通常含有一种隐式煽动，这是一种隐式情感的表达。隐式煽动常以更巧妙的方式表达对特定目标群体的偏见、歧视和仇恨，有时还会减轻表达的程度。在民族和政治事件这一主题的背景下，大多数用户在表达仇恨时会很明确地煽动实施暴力或歧视性的行为，甚至是直接对目标群体进行辱骂。少数较为理性的用户在表达仇恨时会在其评论中隐含一定的煽动性，利用讽刺性的语言和基于刻板印象的言论尽可能地降低自身言论可能带来的风险和舆论的谴责。

此外，从计算语言学和自然语言处理的角度来看，当我们利用语料库进行仇恨言论自动检测研究时，所有的这些标注项都将具有非常重要的参考价值。

5.2 标注方案讨论

由于在前文中我们描述的标注方案具有较高的复杂性，并且仇恨言论这一主题会涉及主观性，在专家标注工作（A组）结束后，我们还随机邀请了以俄语为母语的留学生对已标注的文本（1220条）进行二次标注（B组），通过对比实验和数据分析来应对标注工作的复杂性和主观性。

两组（A组和B组）的标注结果对照情况如表2所示：

	仇恨言论	**攻击性**	**冒犯性**	**讽刺性**	**刻板印象**	**辱骂性**
专家	0.38	0.75	0.88	0.13	0.11	0.86
留学生	0.32	0.68	0.79	0.09	0.08	0.85

Table 2: 两个标注子集中每个标注类别的结果比例，即由专家标注的（表中的第一行）和由本土标注者（以俄语为母语的留学生）标注的结果（第二行）

从表格中我们可以看出，两组标注结果虽略有不同，但总体具有较高的一致性，这证明我们提出的标注方案和规则质量较高，具有一定的可理解性和客观性。我们同时对两组结果的偏差进行了深入分析，认为有以下三个方面：

- **个体差异**。虽然两次标注的方案和规则都是相同的，但由于本土标注者（以俄语为母语的留学生）的母语背景、文化程度和受教育程度不同，导致他们和专家对文本内容、标注规

则的理解不尽相同。尽管标注一致性偏差并不明显，但这些本土标注者的判断并不能与专家的标注结果完全一致。

- **完成质量**。除了标注者的个体差异，标注时的工作态度也会对结果产生影响。我们收到的标注者的反馈表明，他们有的时候可能没有仔细阅读和充分考虑我们给出的标注规则和示例，往往由于自己的疏忽产生了一些不准确的标注结果。

- **规则制定**。两组数据中存在的不一致性表明标注规范中存在一些不足，这些不足会给标注者造成模棱两可的印象。当标注者产生疑惑时，现有规范和示例可能并不能为其提供有效的参考和帮助。此外，四个仇恨程度的级别（1级到4级）之间的区别通常基于语用而非语义特征，这将导致标注者更加重视评论者的态度，而不是其评论文本的实际内容。

此外，通过对比以上两组数据我们发现，专家标注者之间存在的最大分歧在于仇恨的强度。例如，根据我们的标注方案和规则，表1中的评论文本（7）在仇恨程度上应当被认为比文本（8）更为强烈、更加危险，仅仅是因为前者的评论者使用的是第一人称的语句结构，会涉及到个人的责任，而后者使则用了更独立的形式，没有涉及到评论者本人。但另一种观点认为，文本（8）的仇恨程度可能高于文本（7），由此产生了分歧。相反地，本土标注者（以俄语为母语的留学生）在仇恨程度方面的标注具有高度的一致性。本土标注者对于仇恨言论的标注一致性最高，对仇恨程度的标注一致性次之。这说明本土标注者之间具有较高的标注一致性和可靠性。当专家和本土母语标注者出现分歧时，可以优先考虑本土标注者的标注结果。

鉴于以上对于标注结果的观察，我们发现仇恨程度这个系数在我们的方案中是最有争议的一项。我们的研究结果表明，并非所有仇恨言论的表达方式都是相同的，而且仇恨的强弱程度和表达色彩也不尽相同。在有效定义仇恨强弱程度和表达色彩之前，仍有许多工作要做。

我们制定的标注方案和规则仍然存在一些不足，不能够解决在标注时遇到的所有问题。因此，日后我们对标注方案的思考和制定还有很大的提升空间，尤其是在仇恨强度的概念厘定以及对其进行标注的方式上仍需要改进，需要制定出一个更简单、更具有普适性的方案，例如在标注仇恨言论的程度时只标注为“强”或“弱”(Del Vigna12 et al., 2017)，或是对攻击性和冒犯性在表达色彩方面提出更加清晰和细致的规定等等。

6 结论

在本文中，我们构建了针对乌克兰民族和政治问题的俄语网络仇恨言论语料库，并提出了一种全新的、多维度的语料标注方案，以更深入地研究仇恨言论这一颇具复杂性的问题。具体而言，我们对俄语社交网络评论文本进行了基于话题的选取和收集。除了对评论文本中是否含有仇恨言论进行判断，我们还标注了它的强弱程度（1级到4级），以及攻击性、冒犯性、讽刺性、刻板印象和辱骂性的存在与否和强弱程度。

我们对标注的结果进行了初步分析，总结了仇恨言论的具体特征，同时发现这样层次丰富且细粒度的标注方案并非没有缺陷，这些问题都已在本文示例中予以指出和讨论。该研究一方面为俄语网络仇恨言论的研究开辟了新的前景，尤其是为俄语网络仇恨言论自动检测方法提供了数据基础；另一方面，由于语料库自身的复杂性，我们所构建的语料库虽为基于话题的网络仇恨言论语料库，但我们认为，它不仅可以为仇恨言论本身的研究提供基础，还可能更详细、更系统地用于对其他语言现象的分析和研究。

致谢

感谢所有匿名评审人对本文的审阅，感谢首都师范大学隋然教授、王宗琥教授、北京外国语大学武瑗华教授、北京大学王辛夷教授对本文的建议与启发，谢谢！

参考文献

Azalden Alakrot, Liam Murray, and Nikola S Nikolov. 2018. Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181.

Ali Alshehri, Hassan Alhuzali El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2018. Think before your click: Data and models for adult content in arabic twitter. In *TA-COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety*, volume 15.

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The maneco corpus and some input from critical discourse analysis. *arXiv preprint arXiv:2008.06222*.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Karmen Erjavec and Melita Poler Kovačič. 2012. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Filip Klubička and Raquel Fernandez. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. *arXiv preprint arXiv:1805.04661*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Alessandro Maisto, Serena Pelosi, Simonetta Vietri, Pierluigi Vitale, and Via Giovanni Paolo II. 2017. Mining offensive language on social media. *CLiC-it 2017 11-12 December 2017, Rome*, page 252.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Max Weber. 2009. *The theory of social and economic organization*. Simon and Schuster.

Екатерина Николаевна Василенко. 2019. Язык вражды в заголовках новостных интернет-статей (на материале белорусского сегмента всемирной сети). *Вестник Курганского государственного университета*, (1 (52)):80–84.

Павел Николаевич Хроменков. 2016. Лексика вражды в публичной политической риторике периода холодной войны (на материалах инаугурационных речей президентов США середины xx в.-80-х гг. xx в.). *Вестник Московского государственного областного университета. Серия: Лингвистика*, (3):107–117.

ММ Шарнин, НС Ищенко, and НЮ Пахмутова. 2018. Использование методов тематического моделирования многоязычных коллекций для прогноза тревожных событий. In *Шестнадцатая Национальная конференция по искусственному интеллекту с международным участием КИИ-2018*, pages 297–304.