《中文信息学报》稿件排版格式

文章编号：1003-0077（2011）00-0000-00

基于语法的维吾尔语情感词汇自动获取

玛尔哈巴·艾赛提[1]，艾孜尔古丽[1]，玉素甫·艾白都拉*

（新疆师范大学 计算机科学技术学院，新疆 乌鲁木齐 830054）

摘要：情感词汇的获取是文本倾向性分析的基础。为了解决人工识别方法低效的不足，并为维吾尔语情感词的研究及情感词词典的创建提供一些可供选择的方法和思路，本文首先分析了维吾尔语情感词汇在上下文中表现的特征，并结合维吾尔语本身的语法特征，建立了扩展的维吾尔语新增特征模型，与词频逆文档频率(TF－IDF)算法相结合，实现了维吾尔语情感词汇的识别。实验结果指出此特征模型有效地提高了情感词汇的识别率。

关键词：情感词汇，维吾尔语；语法；自动获取

中图分类号：TP391　　**文献标识码**：A

Automatic acquisition of emotional words in Uyghur language based on grammar

Merhaba Eset[1], Azragul[1] ,Yusup Abaydulla*

(School of Computer Science & Technology . Urumqi , Xinjiang 830054, China)

Abstract: The acquisition of emotional vocabulary is the basis of the analysis of text orientation. To deal with the problem of artificial method is inefficient, and for selection of methods and ideas which are created to provide some available for dictionary which is not only emotional words but sentiment words in Uyghur as well. Firstly, the text analyzes the traits of Uyghur emotional words in context, and combined with the grammatical features of Uighur, established extensional NEW characteristic model of Uyghur, and with frequency inverse document frequency (TF-IDF) algorithm are combined to achieve the recognition of Uyghur emotional words. The experimental results show that this feature model can effectively improve the recognition rate of emotion words.

Key words: emotional words, Uyghur ; grammar ; automatic acquisition

1 引言

* **收稿日期：2016-08-5**　　**定稿日期：**

基金项目：国家自然科学基金项目（项目编号：61262066，61662081）；国家社科基金重点项目（项目编号：14AZD11）；国家语委重点项目（项目编号：ZD135-28）；新疆维吾尔自治区自然科学基金（项目编号：2014211A045）；新疆维吾尔自治区哲学社会科学研究规划基金项目（项目编号：14CYY093）；教育部人文社会科学一般项目（项目编号：14YJC740001）；国家自然科学基金重点项目（项目编号：61132009；国家自然科学基金（61163064）；教育部人文社会科学工程科技人才培养专项（15JDGC022）；2015-2016年度新疆师范大学文学院研究生创新基金项目（ZYW2015005）；国家少数民族语言资源监测中心项目。

作者简介：玛尔哈巴·艾赛提（1）（1986——），女，硕士研究生，研究方向：计算语言学；艾孜尔古丽（1）（1987——），女，博士，主要研究方向：计算语言学、自然语言处理；通信作者（*）：玉素甫.艾白都拉（1958——），男，教授，主要研究方向：计算语言学、自然语言处理。

随着互联网的普及，各类维吾尔语网站不断地建立并成为维吾尔族网民学习、聊天、讨论、争议的主要平台。这些平台内容包含大量的带有感情色彩的言语及评论，而这些内容对政府了解民情、掌握舆论导向，对企业了解客户对产品的反馈等方面具有重大的现实意义。

维吾尔语情感词的获取并创建情感词典是识别文本情感倾向性的基础，为网络舆情分析和网络内容安全提供基础性资源。本文以维吾尔语语法特点为基础，首先分析了维吾尔语情感词汇在上下文中表现的特征，建立了扩展的维吾尔语新增特征模型，设计算法，提高情感词汇自动获取的准确率。

2 相关工作

文献[1]指出，情感分析技术可分为两类，分别是基于机器学习的方法和基于情感词典的方法。前者通过大量的主观语料，分析情感词汇的语境信息，抽出特征，再进行情感分析。如：文献[2]中，将情感词汇的语法规律与 CRF 模型相结合，实现了中文情感词汇的自动识别。文献[3]，分析了维吾尔语情感词汇的语境特征，建立特征模板，再利用条件随机场模型实现维吾尔语情感词汇的自动获取。基于情感词典的方法，主要思想是根据已建立好的情感词典所提供的语义关系来判断文本的倾向性。如：王志涛等[4]利用统计和点间互信息识别新的情感词，构建新情感词词典，提出了基于词典和规则集的中文微博情感分析方法。年梅等[5]首先，创建维文情感种子词集，并利用同义词词典扩展；其次对 HowNet、NTUSD 以及大连理工大学开发的情感词典进行并运算，翻译为维吾尔语词汇构成候选词集合；最后计算候选词与种子词之间的点互信息值，判别极性。

目前，国外在此领域已有了显著的成果，如：比较著名的英文情感词词典，WordNet、General Inquirer(GI)、Opinion Lexicon 等。在国内，虽然中文情感分析的研究相对国外起步较晚，但无论是理论还是应用上都有了一定的成果。如：HowNet 情感词语集、NTUSD 以及大连理工大学开发的情感词典等。而维吾尔语情感分析研究刚起步不久，相关资料极少。到目前为止，还没有公开共享的维吾尔语情感词库或情感词词典可以下载并使用。在传统语言学研究领域中，对于维吾尔语情感词汇这个概念只有中央民族大学少数民族语言文学博士阿布都鲁甫·塔克拉玛干尼教授在他的“维吾尔语词汇学与研究”书上[6]提及，他把维吾尔语词汇分为两大类：基本词汇和一般词汇；其中一般词汇包括情感词汇，并维吾尔语命名为：هېس-تۇيغۇلۇق سۆزلەر。查阅近年有关文献可知[7-10]，维吾尔语句子、文本级别情感分析研究较多，词汇级别的研究比较少。情感词汇的识别需要耗费大量的人力和时间，自动获取情感词汇可以解决人工获取低效的缺点。对词汇进行倾向判断过程，基本等价于情感词典的构建，也是文本情感分析的基础。

本文分析、研究维吾尔语情感词汇语境信息，并结合维吾尔语本身的语法特征，建立了较完善的维吾尔语情感词汇特征模型，并与词频逆文档频率(TF－IDF)算法相结合实现了维吾尔语情感词汇的识别。

3 情感词汇语法特征

文献[3,11]中总结的维吾尔语情感词汇特征有：词性规律、词和词性的搭配规律、副词修饰规律、否定词搭配规律及连词规律。这些语言特征在各语种中通用，很少涉及到维吾尔语本有的语言特点。文本基于维吾尔语语法特点，在以上基础上，分析维吾尔语情感词汇及上下文特征，总结了以下维吾尔语情感词汇规律：

3．1 情感感叹词规律

根据词性划分的要求，现代维吾尔语词汇可划分为 12 大类。与汉语相同[12]，维吾尔语情感词汇主要分布在形容词、名词、动词及副词中，在各个词性的分布并不均匀，这种不均匀性符合分类特征选取的标准。

除此之外，维吾尔语中经常使用感叹词来表达强烈的情感。感叹词是本身没有特定词汇意义而独立于句中的其他成分，表达感情色彩、赞成、呼叫、答应等附加意义的语类。根据附加意义感叹词可分为情感感叹词、应答感叹词和呼叫感叹词三种类型[13]。对于句子的情感有直接影响的是表示人类喜、怒、哀、乐等心理活动的情感感叹词。因此句中出现情感感叹词可以断定此句为情感句，并将句中出现的形容词识别为情感词。如表 1 所示。

表1　维吾尔语情感感叹词分类

倾向性	注释	
褒义	表示满意、赞扬的情绪	ھەبەللى،بەللى،بەللى-بەللى ،باركاللا...
	表示兴奋、惊喜的情绪	پاھ،ۋاھ،ئوھۇي،ئوھۇ...
贬义	表示惋惜、厌恶、失望的情绪	ھەي،ئاپلا،.ئەستاغپۇرۇللا،ئىسست،ۋاي -ۋۇي...
	表示惊讶，祈求等心情	توۋا،يائاللا...

3．2 情感词附加成分规律

维吾尔语是黏着语，这种黏着性表现在用附加成分和词缀来实现一些语法功能。维吾尔语附加成分大致分为两类，构词附加成分和构型附加成分。构词附加成分追加在词干后形成新的派生词。情感词也包含大量派生词，派生词由词根和构词词缀两部分组成，部分构词词缀的追加能够影响甚至确定情感词的倾向性。本文将情感词的附加成分总结为两类，一是追加后直接影响情感词倾向性的词缀，另外一种是常跟情感词搭配出现的附加成分。

3.2.1 影响情感词极性的词缀

形容词是情感词分布最多的词性。本文分析构建派生形容词的附加成分，统计出以下直接影响情感词倾向性的词缀。如表 2 所示：

表2　影响情感词极性的词缀

举例	词性	倾向性	注释	词缀
باتۇرانە（勇敢的） ئەمگەكچان（勤奋的） مەرىپەتپەرۋەر（提倡教育的） ئىقتسادچىل （勤俭节约的）	形容词	褒义词	后缀，构建热爱、赞美、敬佩该名词所表达的事物特征的形容词	ـانە，ـچان ， ـپەرۋەر，ـچىل
ئاماخور（贪婪的 ） شۇھرەتپەرەس（贪图功名的）	形容词	贬义词	后缀，追加此词缀形成的词表示过分的追求某件事	ـپەرەس，ـخور
بەتنىيەت（坏心肠） ئادىنساپ（没良心） بىھۆرمەت（不尊重的）	形容词	贬义词	前缀，构建不具备或否认该名词所表达事物特征的形容词	بەت- نا- ，بى-
يۈزسىز（不要脸的） ساپاسىز（没素质）	形容词	贬义词	后缀，构建表示该名词所表达的事物的不存在或者非常少的形容词	ـسىز

3.2.2　情感词缀搭配

a.形似格词缀“ تەك\ دەك ”与情感词的关系

维吾尔语中的形似格词缀“ تەك\ دەك ”加在一些词汇末尾，表示人或事物相互间的性质、特征等方面具有可比喻的共性。此词缀出现的句子中，发言者一般用比喻的手段来表示对某件事或人的态度和看法。因此带有此词缀的词汇前后常出现情感词。如：

شۇڭا مىللەتلەر ئىتتىپاقلىقى ۋە ئىجتىمائىي مۇقىملىقنى كۆز قارىچۇقىمىزنى ئاسرىغاندەك ئاسرىشىمىز كېرەك.

译：“因此，我们应该把民族团结和社会稳定，像爱护我们的眼睛一般进行保护。”句中下划线的词汇含有形似格，形似格后面第一个词汇为情感词ئاسرىشىمىز“保护”。除此之外，“ ئانىدەك كۆكرەك كېرىپ چىقىشتەك تەسىرلىك（母亲般的慈祥）”，“ ھايۋاندەك ياۋۇز（禽兽一样的凶残）”，“

ئىش-پائالىيەتلەر（挺身而出的感人事迹）”等带有形似格的情感词搭配也在语料多处出现。

b.词缀“-لارچە\-لەرچە” 构建的情感词汇

“-لارچە\-لەرچە” 构建带有情感色彩的情态副词，含有此词缀的副词一般表示动作完成的形式和状态，大多数为情感词汇。如：“باتۇرلارچە（英勇地）”，“ۋەھشىلەرچە（残酷地）”，“ئەخمەقلەرچە（愚蠢地）”。

c．带有“-لىق\-لىك\-لۇق\-لۈك”词缀的情感词搭配

“-لىق\-لىك\-لۇق\-لۈك”这是维吾尔语中最有效的构词附加成分之一，追加后可以形成名词和形容词，应用广泛。据分析发现，带有此词缀的词与“بىلەن”后置词和“ۋە”连词共同出现时，会是作为带有情感色彩的抽象名词。如：

كادىرلارنىڭ تۆۋەنگە چۈشكىنى ياخشى ئىش ،لېكىن شەكىلۋازلىق بولۇپ قالسا بولمايدۇ. شەكىلۋازلىق بىلەن ئەپلەپ سەپلەپ كۆننى ئۆتكۈزۈپ كەتسە خەلق ئاممىسىغا ۋە دۆلەتكە يۈز كېلەلمەيدۇ .

译：“干部下基层是件好事，但不能成为形式。以形式主义凑合着过日子，将会对不起民众和国家。”这里的 “شەكىلۋازلىق（形式主义）”与“بىلەن”后置词一同出现，表示“以形式主义”。除此以外，“مۇلايىملىق بىلەن（温柔地）”，“تەشەببۇسكارلىق بىلەن （积极地）”，“ ئالدامچىلىق ۋە خىيالىپەرسلىك（欺骗与空想）”等搭配也常出现在语料中。

3．3 词汇上下文搭配规律

通过情感词的语境信息，某些情感词汇常跟特定的词汇共同出现。此特征可以作为情感词识别的一个标准。

a. “-ھالدا”与情感词汇的搭配

维吾尔语副词“-ھالدا”一般出现在带有情感色彩的词汇后面，表示行为或状态特征。如：

قانۇن – بەلگىلىمىگە خىلاپ ھالدا تېرىلغۇ يەر ۋە ئورمانلىقنى ئىگىلۋېلىش قەتئىي چەكلىنىدۇ.

译：坚决制止违法占有耕地及森林。这里的“-ھالدا”前面的خىلاپ是贬义词“违背”。另，

“ ئۈمۈدسىزلەنگەن ھالدا（绝望地）”， “غەزىپلەنگەن ھالدا（愤怒地）”，“سۆيۈنگەن ھالدا（愉快地）”等搭配在语料中也会多次的出现。

b. 动词“-قىل”形成的情感词汇

维吾尔语动词“-قىل”的原意为“做、弄、搞”。一般与名词、形容词等静词合并构成复合动词，表示词干表达的内容付诸实施的动词。与汉语不同的是，汉语中很多情感词可以是名词，同时也可以作为动词出现，但在维吾尔语中，很多带有情感色彩的形容词或名词后追加助动词，才能形成相对应的动词。

如：①加强财政改革，管理好资金。. مالىيە ئىسلاھاتىنى چوڭقۇرلاشتۇرۇپ،مەبلەغنى ياخشى باشقۇرۇش كېرەك

②我们学校的制度需要改革。. مەكتىپىمىزنىڭ تۈزۈمى ئىسلاھ قىلىنىشى كېرەك

译：在中文，第一句子中“改革”是名词，第二句中 “改革”作为动词出现。但在维吾尔语里第一句中出现的名词“ئىسلاھ”后面追加“-قىل ”才使它转换为动词。维吾尔语中这类现象比较普遍，因此可以利用这类助动词来识别部分情感词。

3．4 词干扩展规律

维吾尔语是黏着语，维吾尔语的词可以分为词干和词缀两个部分，并且一个词干后面可以陆续加上多种词缀。这种现象使维吾尔语同一个词汇在不同的语境里可以扩展为多种样式。比如：情感词سەت（丑）可以扩展为 سەتلىشىش（出丑）、سەتلىمەك（辱骂）、سەتچىلىك（丑事）、سەتلەشمەك（丢人）等新的情感词，并且扩展的新词极性也跟词干一致。

3．5 副词修饰规律

维吾尔语中，修饰情感词的有程度副词，语气副词和情态副词。

a.维吾尔语中的程度副词一般表示动作或事物性质的程度，并修饰带有倾向性的形容词，位置出现在被修饰词前面[14]。如：“بەكمۇ توغرا （非常正确）”，“ تامامەن ئورۇنلۇق（完全合理）”，“سەل ئادىل ئەمەس（有点不公平）”等。

b.除了程度副词外，表示说话者对所发生行为动作的某种语气的语气副词也会修饰情感词。如：“ قەتئىي قارشى（ 坚决反对）”， “چوقۇم رازى（ 一定满意）”等。语气副词后面一般跟的是表示态度的情感词，语气动词往往带有肯定和否定的语气。

c. 情态副词表示动作或性质特征的各种形式和状态，不仅可以修饰情感词，大部分还可以作为情感词汇。

3．6 连词搭配规律

利用词语之间的连词来判断情感词的倾向性是比较常用的方法。早在 1997 年，已提出[15]连词所连接的成分的关联性来判断情感词极性的方法。连词所连接的各成分之间的关系是并列的，也可以是偏正的。由于连词的连接功能是逻辑性的，所以根据连词的逻辑意义及情感词获取的条件，将其归类为并列连词和转折连词。转折连词连接的两个成分通常具有相反的情感倾向。并列连词连接的两个词具有相同的情感倾向。如：

ئورتاق روناق تېپىپ تەرەققىي قىلدىرغان ياخشى ۋەزىيەتنى ھەسسىلەپ قەدىرلىشىمىز ۋە قوغدىشىمىز كېرەك.（来自于新闻评论）

译：“我们应该加倍地珍惜并维护共同发展的大好形式。”句中褒义词“珍惜”和“维护”分别出现在连词“ۋە”的前后，并倾向性一致，这种特点可用于情感词的极性判断。

3．7 否定词搭配规律

维吾尔语中除了否定词以外，还可以通过附加一定的词缀来表示否定[16]。对于情感词的识别有直接影响的是否定词“ئەمەس”,“يوق”，意思与汉语里的“不”、“没”相同。如：

(1) ئۇ ناھايىتى بىلىملىك لېكىن كەمتەر ئەمەس. 他很有学问但不谦虚。

(2) سىزدە ئەيىب يوق. 你没有错。

维吾尔语的否定词一般出现在被修饰词的后面。

4 维吾尔语情感词汇的识别研究

本文根据上述分析的维吾尔语情感词汇特点，将维吾尔语语法规律作为识别的特征。在现有的特征基础下，构建了新增的分析模型。该特征模型更详细地集合了情感词汇的丰富语境信息。下一步，把维吾尔语情感词汇分析模型与带词权重的 TF－ IDF（erm Frequency － Inverse Document Fre-quency）算法相结合来实现情感词汇的获取。

4．1 维吾尔语情感词汇分析模型

情感词汇分析模型的设计是情感词汇获取的关键，同时非常依赖于情感词汇本身的特征或规则，情感词汇规律获取的质量直接影响着识别结果。本文在文献[3,11]中常用的维吾尔语情感词一般规律的基础下扩展新增了感叹词搭配特征、附加成分搭配特征、词汇上下文搭配特征及词干扩展特征。表 3 是根据文献[3,11]中已提出的情感词一般规律所建的分析模型（标为情感词汇分析模型 1），表 4 是本文扩展新增的特征模型。

表 3　情 感 词 汇 分 析 模 型 1

特征名称	符号表示	说明
词性特征	pos（0）	情感词词性
	pos（-1）	情感词前面第一个词词性
	pos（1）	情感词后一个词词性
词性搭配特征	pos（-1）pos（0）	情感词与前一个词词性
	pos（0）pos（1）	情感词与后一个词词性
	pos（-2）pos（0）	情感词与前第二个词词性
	pos（0）pos（2）	情感词与后第二个词词性
副词修饰特征	adv(-1)	情感词前面第一个词是否是副词
	pos(0)adv(-1)	情感词词性与前一个是否是副词
否定词特征	prv（1）	情感词后一个词是否是否定词
	pos（0）prv（2）	情感词词性与后面第二个词是否是否定词

表 4　新 增 情 感 词 汇 分 析 模 型

特征集	特征模板	说明
感叹词搭配特征	jec(1) jec(2) jec(3)	情感词后面的第 1、第 2 、第 3 个词是否是感叹词

	jec(-1) jec(-2) jec(-3)	情感词前面的第 1、第 2、第 3 个词是否是感叹词
附加成分搭配特征	afx_1(-1)	情感词前面第一个词是否带词缀“دىك،تىك”
	afx_1(-1) pos(0)	情感词词性和前面第一个词是否带词缀“دىك،تىك”
	afx_2 （0）pos（0）	情感词词性与情感词是否带词缀 “لىق،لىك،لۇق،لۈك”
	afx_3（0）	情感词是否带词缀 “لارچە،لەرچە”
词汇上下文搭配特征	pos（0）word(1)	情感词词性和后面第一个词是否是“قىل”或“هالدا”
词干扩展特征	stem（0）	候选词词干是否是情感词

4．2 维吾尔语情感词汇识别过程

实验语料来自于国家语言资源监测中心少数民族分中心维吾尔语文研究基地提供的已标注的维吾尔语小学语文教材。实验采用 4 倍交叉验证(4-fold crossvalidation) ，即将语料随机分为四份，其中三份做训练集，一份做测试集。识别步骤如下：

首先，将训练语料进行分词处理，再对已建立的特征模型进行训练，根据训练的效果对每一个特征集的作用给予权值。接下来，同样对测试语料进行分词处理，依次输入到识别模块中,根据训练集所保存的权值进行识别,权值匹配将给识别标签赋值为 true,相反则 false。

4．3 带词权重的 TF－ IDF 算法

TF－IDF （ term Frequency－Inverse Document Frequency) 是一种统计方法，语料库中某一文件内的高词语频率（TF)，以及该词在整个文件集合中的低文件频率(IDF)，可以产生高权重的 TF－IDF 值。因此，该算法倾向于过滤掉常见的词语，保留重要的词语。利用该算法在一定的程度上过滤掉维吾尔语文本中的频率较高但不是关键的词汇，如停用词。这样可以有效地降低非关键词的干扰。TF－IDF 公式如下：

$$W_{tf\text{-}idf} = tf_i(D) * \log(Dw/D_i)$$

其中，$tf_i(D)$为该词在文章 D 中出现的频率，由该词在文档 D 中出现的次数除以文档 D 中所有词出现的次数之和所得。Dw 为所有文档数，D_i 为包含该词的文档数。

5 试验和分析

5．1 实验过程

为了验证新增情感词汇分析模型对维吾尔语情感词汇识别的作用，实验先只使用情感词汇分析模型 1，来识别维吾尔语情感词汇。接下来，依次加入新增分析模型的特征集结果进行对比。

步骤 1：对情感词汇分析模型 1 进行识别；

步骤 2：计算识别的情感词汇的 TF-IDF 值；

步骤 3：在步骤 1 的基础上，依次加入新增情感词汇分析模型中的特征集，每一个特征集加入后重复一次步骤 2，消除假情感词；

新增特征集 1：加入感叹词搭配特征；

新增特征集 2：特征集 1 的基础上加入附加成分搭配特征；

新增特征集 3：特征集 2 的基础上加入词汇上下文搭配特征；

新增特征集 4：特征集 3 的基础上加入词干扩展特征；

步骤 4：分别计算分析模型 1 和 4 种新增特征情况下的正确率、召回率和 F 值；

步骤 5：对比情感词汇分析模型 1 和新增情感词汇分析模型的结果是否有提高。

本文采用计算准确率 P，召回率 R，F-meaure 值 FB1 来评测实验结果。计算方法如下：

$$P = A / B * 100\ \%;$$

其中，A 表示自动判断结果中，判断正确的情感词数；B 表示所有的自动判断为情感词汇的总数；P 衡量的是识别方法的查准率。

$$R = A / C * 100\ \%;$$

其中，C 表示人工标注中的情感词汇总数；衡量的是检索系统的查全率。

$$FB1 = (2 * P * R / P + R) * 100\% ;$$

5．2 实验结果

实验从语料中随机筛选 168 篇文章，包含 15877 个词汇，通过扩展的新增特征模板识别的情感汇有 5612 个，经过过滤和去重等操作最终获得的正确的情感词有 2834。实验结果表 5 所示：

表 5　实 验 结 果 （%）

特征集		正确率（%）	召回率（%）	F1 值
情感词汇分析模型 1		63.1%	47.4%	54.1%
新增情感词汇分析模型	新增特征集 1	65.3%	51.2%	57.3%
	新增特征集 2	68.1%	55.6%	61.2%
	新增特征集 3	69.3%	54.6%	61.0%
	新增特征集 4	70.2%	56.2%	62.4%

5．3 实验结果分析

结果表明，利用情感词汇分析模型 1 来获取的情感词汇正确率为 63.1%。随着新增情感词汇分析模型的加入，识别率最终达到了 70.2%。实验说明，将语言本身的语法规律作为识别特征并构建为分析模块，能更详细地集合了情感词汇的丰富语境信息，提高了情感词识别的准确性。表 6 将进一步分析每一个新增特征产生的效能差异。

表 6　实 验 结 果 分 析

特征名称	识别结果	结果分析
感叹词搭配特征	خاتىرجەم،گاڭۋاڭ،غېرىب رەتسىزگارماڭ،پەخرى.. （平顺的、糊涂、孤独、多彩、自豪。。。）	增加此特征后正确率提高了 2.2%、召回率提高了 3.8%。此特征有效地召回了对程度副词、连词、否定词没有修饰关系的一些情感词汇。如： باريكاللا/x ، نەمدى/c سەندىن/d P/ خاتىرجەم a/ بولدۇم/v . 句中出现情感感叹词“（赞）باريكاللا”，因此确定此句为情感句，选出其中的形容词“（安心的）خاتىرجەم”。
附加成分搭配特征	سەۋرچان،نىتاقچان، نامەرد،غەلىبلىك،زۇلمەتلىك، بىمەنە،قارىغۇلارچە، غالجىلارچە.. … （耐心的、服服帖帖的、不仁义的、成功的、黑暗的、荒唐、盲目地、猖狂地。。。）	此特征根据维吾尔语的派生构词原理，分析出追加后形成带有情感色彩的词汇的词缀，并用匹配的方式来识别。因此，此特征集不依赖于上下文，避免一些特殊语境信息的干扰。如，用“ لارچە،لەرچە”词缀所识别的词汇有， ”قەھرىمانلارچە،سېخىيلارچە،يۇۋاشلارچە،سەبىيلەرچە،باتۇرلارچە ،ئۆزلارچە ،قانداقلارچە....“ （英勇地、慷慨的、野蛮的、单纯的、勇敢地、以他们的。。。、怎么就。。。。） 下划线的词汇是识别出来的带有相同词缀的假情感词，通过计算 TF－IDF 值，可以有效的排除这类干扰词。增加此特征集后正确率和召回率提高了 3%左右。
词汇上下文搭配特征	（维护）ھىمايە قىلىپ （违反）خىلاپلىق قىلغان （污蔑）تۆھمەت قىلىش （痛苦地）ئازابلانغان ھالدا （得意的）كۆرەڭلىگەن ھالدا	识别结果表示，通过此特征集识别的主要是带有情感色彩的动词。动词是维吾尔语中较复杂的范畴，其形态变化丰富多样。复合动词在情感词中也有一定的分量，而两个复合成分中间以空格来分离。维吾尔语分词也是鉴于词之间的空格来实现的，此特征集可以部分解决将复合动词分为多个词来处理而导致的识别率下降问题。
词干扩展特征	ۋەيران،خانىۋەيران،ۋەيرانە، ۋەيرانچىلىق... （毁坏的，败落，坍塌，荒废。。。）	以上特征是根据维吾尔语中有规律的形态变化和搭配原理来实现识别的。而一个词干可以追加多个词缀，使得维吾尔语词汇的形态变化变得复杂。因此，根据词干获取特征可以避免扩展词汇导致的数据稀疏现象，使召回率提高了 1.6%。

虽然，这种方法提高了情感词汇获取的准确率，然而与国内外汉语、英语等语言相比，

识别率还是较低。本文使用的是人工词性标注的语料，因精力有限词性标注有可能做不到绝对的正确。词性标注的不完善性可能影响情感词汇识别率。加之维吾尔语的形态变化多样，语法结构复杂，对维吾尔语情感词汇的规律分析，特征发现的问题上还有待完善和提高的空间。

6 结论

本文从维吾尔语的语法特点出发，分析了维吾尔语情感词汇在上下文中表现的特征，在此基础上，建立了扩展的维吾尔语新增特征模型，并与词频逆文档频率(TF－IDF)算法相结合实现了维吾尔语情感词汇的自动识别。与此同时，对维吾尔语情感词汇的特征以语言的语法角度做了比较详细的总结。实验结果验证了该方法有效地提高了维吾尔语情感词汇的识别，降低了人工获取的工作量。本次研究是维吾尔语情感词汇获取的初步探索，下一步在此基础下继续钻研，实现维吾尔语文本的倾向性识别。

参考文献

[1]王科,夏睿.情感词典自动构建方法综述[J].自动化学报,2016,04:495-511.
[2]陈建美,林鸿飞,杨志豪.基于语法的情感词汇自动获取[J].智能系统学报,2009,02:100-106.
[3]冯冠军,禹龙,田生伟.基于 CRFs 自动构建维吾尔语情感词语料库[J].现代图书情报技术,2011,03:17-21.
[4]王志涛,於志文,郭斌,路新江.基于词典和规则集的中文微博情感分析[J].计算机工程与应用,2015,08:218-225.
[5]年梅,范祖奎,刘若兰.维吾尔语褒贬情感词典构建研究[J].计算机工程与应用,,:1-5.
[6]阿布都鲁甫·塔克拉玛干尼. 维吾尔语词汇学与研究[M]. 北京：民族出版社，2011.2；41-46.
[7]黄俊,田生伟,禹龙,冯冠军.基于维吾尔语情感词的句子情感分析[J].计算机工程,2012,09:183-185.
[8]黄俊.维吾尔语文本情感分析研究[D].新疆大学,2013
[9]热依莱木·帕尔哈提,孟祥涛,艾斯卡尔·艾木都拉.基于区分性关键词模型的维吾尔文本情感分类[J].计算机工程,2014,10:132-136+142.
[10]罗亚伟,田生伟,禹龙,吐尔根·依布拉音,艾斯卡尔·艾木都拉.意见挖掘中维吾尔语文本隐式情感分析[J].计算机工程与设计,2014,09:3295-3300.
[11]禹龙,田生伟,冯冠军.维吾尔语情感词汇自动识别[J].计算机工程,2011,07:213-215.
[12]周晓. 基于互联网的情感词库扩展与优化研究[J]. 东北大学信息科学与工程学院,2011.06；14-15；
[13]力提甫·托乎提.现在维吾尔语参考语法[M]. 中国社会科学出版社，2012；206-213.
[14]热孜亚木·麦麦提吐逊，买提热依木·沙依提.汉语-维吾尔语副词对比研究[J].中央民族大学 维吾尔语言文学系，2012.4,；18-34.
[15] Hatzivassiloglou V，McKeown K R． Predi－cting the Semantic O-rientation of Adjectives[C].Proc. of the 35th Annual Meeting of the Association for Computational Linguistics. Madrid，Spain: [s. n.]，1997: 174 －181.
[16]王海蓉,孙丽莉. 汉维语双重否定语形对比[J]. 塔里木大学学报,2010,01:77-83.

作者联系方式:玛尔哈巴·艾赛提　新疆乌鲁木齐沙依巴克区友好南路 91 号 9-2-1002 830000
15999107827　278416557@qq.com

作者简介：

玛尔哈巴·艾赛提（1986——），硕士研究生，主要研究领域为计算语言学、自然语言处理。

Email:278416557@qq.com；

艾孜尔古丽·玉素甫（1987——），讲师，主要研究领域为计算语言学、自然语言处理。

Email:Azragul2010@126.com；

通讯作者：玉素甫·艾白都拉（1958——），教授，主要研究领域为计算语言学、自然语言处理。

Email:ysp2002@126.com；