计算机科学
COMPUTER SCIENCE

# 基于大规模用户视频弹幕的颜文字自动化发现

毛馨，雷瞻遥，戚正伟

引用本文

毛馨，雷瞻遥，戚正伟. 基于大规模用户视频弹幕的颜文字自动化发现[J]. 计算机科学，2024，51(1)：284-294.

MAO Xin, LEI Zhanyao, QI Zhengwei. Automated Kaomoji Extraction Based on Large-scale Danmaku Texts [J]. Computer Science, 2024, 51(1): 284-294.

---

**相似文章推荐（请使用火狐或 IE 浏览器查看文章）**

**Similar articles recommended (Please use Firefox or IE to view the article)**

无监督句对齐综述

Survey of Unsupervised Sentence Alignment

计算机科学, 2024, 51(1): 60-67. https://doi.org/10.11896/jsjkx.231100024

面向国产深度学习平台的自然语言处理模型迁移研究

Study on Model Migration of Natural Language Processing for Domestic Deep Learning Platform

计算机科学, 2024, 51(1): 50-59. https://doi.org/10.11896/jsjkx.230600051

SemFA:基于语义特征与关联注意力的大规模多标签文本分类模型

SemFA:Extreme Multi-label Text Classification Model Based on Semantic Features and Association Attention

计算机科学, 2023, 50(12): 270-278. https://doi.org/10.11896/jsjkx.230300239

基于可信细粒度对齐的多模态方面级情感分析

Aspect-based Multimodal Sentiment Analysis Based on Trusted Fine-grained Alignment

计算机科学, 2023, 50(12): 246-254. https://doi.org/10.11896/jsjkx.221100038

多层面语义结构增强的对话情感诱因片段抽取

Multi-level Semantic Structure Enhanced Emotional Cause Span Extraction in Conversations

计算机科学, 2023, 50(12): 236-245. https://doi.org/10.11896/jsjkx.221100189

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400120

# 基于大规模用户视频弹幕的颜文字自动化发现

毛 馨 雷瞻遥 戚正伟
上海交通大学电子信息与电气工程学院 上海 200240
(maoxin@sjtu.edu.cn)

**摘 要** 作为网络时代产生的新型表情符号,颜文字不仅受到了网络用户与社会主流媒体的青睐,被广泛应用于网络文本中,而且在情感表达、文化宣传等方面具有独特的价值。鉴于颜文字具有丰富的语义情感信息,结合颜文字对网络文本进行研究,能够促进对网络文本的分析与理解,提高多项自然语言处理任务的效果。对文本中的颜文字进行检测与提取,是结合颜文字进行文本分析的首要步骤;然而,由于颜文字具有结构灵活、种类丰富、更新换代快等特点,现有工作大多缺乏对颜文字的整体分析,具有准确率低、边界确定困难、时效性差等局限性。文中通过深入分析颜文字的特征,提出了一种基于大规模弹幕文本的颜文字检测与提取算法 Emoly。该算法通过预处理方法提取出初步候选字符串,将多种改进的统计指标与过滤规则相结合,用于筛选出最终候选字符串,并通过文本相似度对其排序,输出最终结果。实验结果表明,Emoly 算法在百万规模的弹幕文本中达到了 91%的召回率,能够全面而准确地将文本中的颜文字检测并提取出来,具有稳健性、优越性与通用性。同时,该算法还为中文分词、情感分析、输入法词库更新等任务提供了新的解决思路与方法,具有广泛的应用价值。

**关键词**:自然语言处理;数据分析;颜文字;视频弹幕
**中图分类号** TP391

## Automated Kaomoji Extraction Based on Large-scale Danmaku Texts

MAO Xin, LEI Zhanyao and QI Zhengwei
School of Electronics, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

**Abstract** As a new type of emoticon symbol that emerged in the Internet age, kaomoji not only enjoys popularity among Internet users and mainstream social media but also has indispensable value in emotional expression, cultural promotion, and other aspects. Considering that kaomoji carries rich semantic and emotional information, studying them in the context of Internet texts can promote the analysis and understanding of such texts, thus improving the effectiveness of various natural language processing tasks. Detecting and extracting kaomoji from texts are the primary steps in analyzing texts with kaomoji. However, due to the flexible structure, diverse types, and rapid evolution of kaomoji, most existing works lack a comprehensive analysis of kaomoji, resulting in limitations such as low accuracy, difficulty in determining boundaries, and poor timeliness. In this paper, through an in-depth analysis of kaomoji features, a kaomoji detection and extraction algorithm called Emoly based on a large-scale danmaku text dataset is proposed. It extracts preliminary candidate strings through preprocessing methods, combines various improved statistical indicators and filtering rules to select the final candidate strings, and ranks them based on text similarity to produce the final results. Experimental results show that the Emoly algorithm achieves a recall rate of 91% in a dataset of millions of danmaku texts, effectively and accurately detecte and extracte kaomoji from the texts. It demonstrates robustness, superiority, and generality. Additionally, the proposed algorithm provides new ideas and methods for tasks such as Chinese word segmentation, sentiment analysis, and input method dictionary updates, offering broad application value.

**Keywords** Natural language processing, Data analysis, Kaomoji, Video danmaku

## 1 引言

据第 49 次《中国互联网络发展状况统计报告》,截至 2021 年 12 月,我国互联网用户数量达到了 10.32 亿,互联网普及率由 2017 年的 55.8%飞升至 73%[1]。弹幕便是网络语言文化所衍生出的典型代表。

“弹幕”一词源于军事领域,本意是用密集炮火对某一区域进行轰击[2]。2006 年,日本视频网站 Niconico 动画首先将视频评论以流动的方式实时显示在画面上。在此功能下,实时评论会像子弹一样从视频屏幕中穿过,因此这种评论显示

到稿日期:2023-04-17 返修日期:2023-10-31
基金项目:国家自然科学基金(62141218)
This work was supported by the National Natural Science Foundation of China(62141218).
通信作者:戚正伟(qizhwei@sjtu.edu.cn)

效果被称为“弹幕”。发展到现在，弹幕一词已不仅仅用于指代这一显示效果，而是泛指这种效果下的评论本身。在一些研究工作中，弹幕也被称为视频实时评论（Time-Sync Comments）[3-5]。

弹幕所蕴含的用户意见、情感评价等信息，在视频评价、个性化推荐等多项研究领域展现出了极大的价值。近年来，不仅越来越多的视频与社交平台引进弹幕功能，也有越来越多的研究人员开始从不同方向对弹幕进行研究[6-9]。

“颜文字”是弹幕文本的重要组成部分之一，其是指由多个字符构成的，形似人脸表情、肢体动作或物体的一组字符组合。“：－）”是一个表示“笑脸”的颜文字，该颜文字被认为最早由卡内基·梅隆大学的斯科特·法尔曼教授发明[10]。

随着弹幕文化的发展，颜文字的影响已经从网络空间延伸到现实社会[11]。首先，颜文字具有传播广、影响力大、受众多的特点。在国内主流社交软件微博中，截至 2022 年 5 月，颜文字话题拥有 1245 万阅读量。其次，颜文字在社会层面也有着愈发广泛的影响力。例如，共青团中央曾以《【走你】山东舰版航母 style ┌(゜ω゜)=☞》[1)] 为名发布视频。除此之外，颜文字还具有巨大的商业价值。淘宝上以颜文字为主题的 T 恤、玩偶等产品以及收录了大量颜文字的各大输入法软件，都体现了颜文字所具有的商业潜力。

已有多项研究工作揭示出颜文字在文本语义传达、情感表达等方面具有不容小觑的作用与重要的研究价值[12-14]。然而，现有文本分析方法大多不适用于对带有颜文字的网络文本进行分析。由于颜文字大多利用视觉特征来传递信息，没有明确的语法定义，并且其组成字符大多是传统意义上的停用词[15]，因此，在传统的文本分析中，颜文字往往被视为停用词而被剔除。这种处理方法会遗漏颜文字的情感与语义信息，从而影响文本的分析结果。颜文字为弹幕文本的处理与分析工作提出了两项挑战，分别是如何从文本中检测并提取出颜文字，以及如何理解颜文字的语义与情感信息。其中，应对第一项挑战是解决第二项挑战的基础，只有将文本中的颜文字正确检测并提取出来，才能对其进行后续研究。因此，本文主要关注大规模弹幕文本中颜文字的检测与提取。

现有的颜文字检测与提取工作[16-20]大多基于词典匹配、规则提取、计算机视觉技术等方法进行，并且在测试集上取得了一定的效果。然而，这些工作皆具有一定的局限性。首先，大多数现有工作过度依赖于现有颜文字，而对新型颜文字的检测效果较差，难以适应复杂多变的颜文字文化。以 Yu 等[16]在 2019 年提出的基于词典匹配思想的颜文字提取方法为例，通过该方法得到了 83.8％的召回率；然而，用本文构建的时间范围在 2018－2021 年的数据集进行实验时，该方法却只达到了 60％的召回率。其次，这些工作大多针对的是标点符号运用规范的文本，对文本的规范程度要求较高，难以适应不规范的弹幕文本。综上，鉴于颜文字的日益广泛使用及其在情感表达、语言分析等方面的重要性与目前颜文字检测技术的局限性之间的矛盾，迫切地需要一种新方法，能够准确、全面、稳定地从大规模弹幕文本中自动化提取颜文字。

针对先前研究工作的不足，本文提出了一种新的颜文字检测与提取算法 Emoly，通过预处理方法提取出初步候选字符串，将多种改进的统计指标与过滤规则相结合用于筛选出最终候选字符串，并通过文本相似度对其排序，得出最终结果，以从大规模弹幕文本中全面、准确地提取颜文字。在自主构建的大规模弹幕数据集上的实验结果表明，Emoly 算法在百万规模的弹幕文本中达到了 91％的召回率，体现了 Emoly 算法的准确性、有效性、稳健性与优越性。除此之外，实验结果还表明 Emoly 算法不仅适用于弹幕文本，而且适用于任意一种网络语言文本。同时，Emoly 算法还为中文分词、情感分析、输入法词库更新等任务提供了新的解决思路与方法，具有广泛的应用价值。本文的主要创新点与贡献如下：

1）针对弹幕文本与颜文字的独有特征（相关分析作为补充材料公开在 figshare 上[21]），本研究将多种统计指标、过滤规则相结合，从而实现有效地对文本中的颜文字与非颜文字字符进行准确分割。

2）创造性地使用文本相似度方法对结果进行评估排序，使算法获得的结果可以更好地服务于后续任务。

3）通过大量用户文本自动挖掘出现有词库未收录的新颖颜文字，形成其他自然语言分析工作能够直接使用、长期更新的颜文字词库，进而将其应用于中文分词、情感分析等任务，解决了因混淆停用词与颜文字或使用的颜文字词库过于陈旧而造成的颜文字情感语义信息缺失的难题，提高了效率与准确度。所提算法检测和提取到的 4 732 个颜文字已全部开源[21]。

本文第 2 章介绍了国内外颜文字的相关工作；第 3 章介绍了基准颜文字集与弹幕数据集的构建；第 4 章介绍了 Emoly 算法的具体实现过程与方法；第 5 章描述了实验过程并对实验结果进行了分析；第 6 章对算法后续应用场景进行了展望；最后对本文工作进行总结。

## 2　相关工作

### 2.1　颜文字定义与分类

颜文字，亦被称为 ASCII 式的表情符号[22]，其与其他字符组合最大的区别在于它的外在视觉特征：其通常形似人脸表情、肢体动作或是植动物等物体；其组成字符往往不再用于其原本用法，而仅仅用于模仿人体或物体的组成部件。以颜文字“(T_T)”为例，该颜文字模仿人脸表情，展现了一张流泪的脸，用于表达悲伤情感。其中，“(”“)”代表人脸，“T”代表流泪的眼睛，“_”代表嘴巴。这些符号都脱离了其原本的用法释义，仅用于模仿人脸中的部件。颜文字对人体物体进行模仿，为文本增添视觉信息，使其更具有画面感[23]。

已有的研究工作大多将颜文字分为两类，分别是西方式颜文字与东方式颜文字[24]。典型的西方式颜文字具有需要倾斜 90°进行阅读的特征，例如字符组合“：－）”及“：－(”（分别表达“笑脸”和“哭脸”含义）就属于西方式颜文字。东方式颜文字无需倾斜方向阅读，并且添加了亚洲文字常见的全角字符，使颜文字可以模仿的对象更加丰富。表 1 列出了部分

1) https://www.bilibili.com/video/BV1yJ411j7kF

颜文字及其类别、模仿对象与释义。

表 1 部分颜文字示例

Table1 Examples of some Kaomojis

| 颜文字 | 类别 | 模仿对象 | 文字释义 |
|---|---|---|---|
| :一) | 西方式 | 人脸表情 | 笑脸 |
| (╥﹏╥) | 东方式 | 人脸表情 | 哭泣 |
| (¦3[▓▓] | 东方式 | 肢体动作 | 睡觉 |
| 〈・))))〉〈〈 | 东方式 | 动物 | 鱼 |

2022 年的一篇研究显示[25]，纯颜文字的弹幕占比大约是 5%。越来越多的研究人员意识到颜文字的价值，开始结合颜文字对网络文本进行研究[26-28]。然而，这类工作大多数都是简单地通过自主收集颜文字，使用词典匹配方法来检测文本中的颜文字[26]，且这种方法收集到的颜文字往往不够全面，无法覆盖文本中全部的颜文字，从而使结果出现误差。本文综合考虑了表 1 中两类颜文字所具有的共同特征，将这两类颜文字皆作为本文工作的研究对象。下面将介绍目前针对这两类颜文字进行的检测与提取的相关工作。

### 2.2 颜文字检测与提取

目前，国内外针对颜文字检测与提取的工作大致分为 3 类，分别是基于组件词典匹配、基于统计和机器学习，以及基于视觉特征的方法。

基于组件词典匹配的方法是对传统词典匹配方法的改进，它通过将颜文字拆分成不同部件，检测文本中是否存在这些部件或其组合，以提取文本中的颜文字。Ptaszynski 等[24]根据人体运动学理论，将颜文字分为左右眼睛、嘴巴、左右边界、左右肢体七大构件，并将左右眼睛与嘴巴视为颜文字的核心构件。他们首先通过互联网收集大量颜文字，构建上述核心构件词典；然后使用构建的词典对文本进行匹配操作，若匹配成功，则将该构件及其左右符号共同构成的字符串视为颜文字提取出来；最后，他们在日本 Yacis 博客语料集中进行了实验，证明了该方法在日文网络文本中的有效性。Yu 等[16]提出的 AZEmo 系统同样使用上述原理对颜文字进行提取，并经由收集自国内视频网站 AcFun、哔哩哔哩动画及购物软件淘宝、天猫的数据集，验证了该方法在中文网络文本中的有效性。这类方法对词典有着强烈的依赖性，其检测与提取流程全部围绕词典进行，并且未对颜文字特性进行系统性分析，虽然能够较好地提取存在于词典中的颜文字，但在处理词典中不存在的颜文字方面相对较弱，难以发现使用新字符创造的颜文字。除此之外，这类方法只以东方式模仿人脸表情的颜文字为研究对象，而未考虑非人脸表情以及西方式颜文字，有一定的局限性。

Tanaka 等[29]的工作是最早在颜文字检测与提取任务上实施机器学习方法的研究之一。他们利用词性类型标签、单词位置等特征，使用 SVM 对颜文字进行检测与提取，主要关注颜文字的词长、字符序列、子串相似度等特征。Kwon 等[30]考虑到了不同颜文字之间具有的相关性，使用 Bi-LSTM 与 Bi-LSTM-CRF 模型对颜文字进行提取。他们使用 BIO 标签，手动对文本中的颜文字进行序列标注并进行模型训练。Yamada 等[18]提出了一种基于字符 N-gram 统计的颜文字提取方法。上述研究虽然从统计特征与机器学习方面为颜文字的检测与提取提出了新的思路，摆脱了词典匹配的局限性，但缺乏对颜文字所具有的对称性等自身特征的研究，在颜文字的边界字符确定上往往会出现失误。Tanaka 在其研究中提到，以文本“(I) haven't known that. ϕ(_ _)taking notes”为例，该方法错误地提取出了“(_ _)”而非颜文字“ϕ(_ _)”。

近年来，一些研究人员注意到颜文字在视觉上具有的独特特征，尝试利用颜文字区别于其他字符组合的视觉特征来从文本中检测与提取颜文字。Bedrick 等[19]认为，颜文字中特定字符的使用是由它们的图形外观决定的。基于此，他们开发了一套字符图形相似度度量工具，用于衡量字符的字形相似度、镜像关系与字形垂直对称情况。最终，他们根据颜文字的对称视觉特征，将具有字形相似性、镜像关系或是自身对称相似性的字符分组，并成对使用它们来识别颜文字。Yokoi 等[31]除了关注颜文字的对称特征外，还考虑了颜文字中描述眼睛的字符。他们观察到，在收集到的颜文字中，有 93% 的颜文字使用相同或对称的字符表示左右眼睛，因此他们通过定义颜文字眼睛字符规则，基于正则表达式来提取描述眼睛的符号的部分颜文字。基于大多数颜文字的对称结构，他们采用 Damerau-Levenshtein 距离，通过比较眼睛符号两侧字符串的相似度来确定颜文字边界，最终提取出文本中的颜文字。这种方法仅考虑到了具有视觉对称性的颜文字，而难以检测出非对称的颜文字，例如“_(:D)∠)_”；而非对称的颜文字在整体颜文字中占有相当大的比例，包括部分非人脸表情类的东方式颜文字与大部分西方式颜文字。另外，这种方法极易提取出非颜文字的字符串。Bedrick 在其研究中提出，同样拥有对称性质的字符串，例如“..?..”，会对结果造成一定的影响。

## 3 数据集构建

### 3.1 基准颜文字数据集构建

本文从网络上收集颜文字组成颜文字数据集，用于分析颜文字特征。由于收集到的所有颜文字都是经网络大众检验的真实颜文字，因此将其称为基准颜文字。

截至 2022 年 5 月，搜狗输入法[32]在华为应用市场内拥有 34 亿下载量，其中内置了包含中文释义的颜文字，便于我们理解颜文字的含义。因此，本文从搜狗输入法内置颜文字库中搜集了 1144 个颜文字，组成了搜狗基准颜文字数据集(以下简称搜狗颜文字集)。

### 3.2 弹幕数据集构建

本文所使用的弹幕数据集收集自国内知名视频社区哔哩哔哩动画(又称 bilibili 网站，简称 B 站)。B 站是国内最大的视频网站之一，引领着弹幕这种独特的社交潮流。据 B 站官方统计[1),2)]，截至 2021 年底，B 站拥有 2.72 亿月均活跃用户，其中 35 岁及以下月活用户占比超 86%，弹幕累计总数已破 100 亿。本文最终构建的弹幕数据集包括从 2018 年 10 月至

1) https://www.bilibili.com/blackboard/aboutUs.html

2) https://www.bilibili.com/video/BV1tS4y1R72d

2021 年 12 月期间共 400 万条以上弹幕。该数据集具有以下特点：

1）规模大，时间跨度久。弹幕文本达到了百万数量级，时间跨度达到 3 年之久，涵盖超过 10000 个视频。

2）涵盖游戏、生活、鬼畜等多分区类型的视频，弹幕语言风格多样。

3）除了包含中文弹幕外，还涵盖了英文、日文等多国语言的弹幕，具有多语种的特点。

4）拥有丰富及新颖的颜文字。B 站用户年轻化的特点使得其弹幕文本含有大量丰富和新颖的颜文字。

## 4　算法设计与实现

基于上述颜文字特征分析，本文设计了一种面向大规模弹幕文本的颜文字检测与提取方法，记为 Emoly，图 1 给出了其完整的算法流程。本节将以图 1 为基础，对 Emoly 算法的设计思路与实现进行具体阐述。

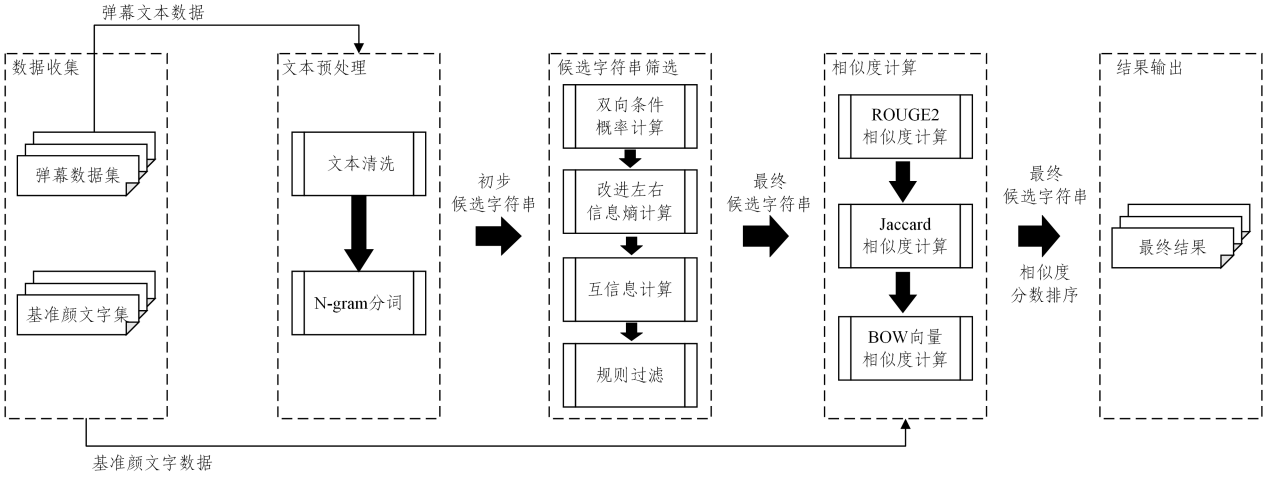



图 1　Emoly 的流程图

Fig. 1　Emoly workflow

### 4.1　候选字符串生成：基于颜文字特征的文本预处理

文本预处理包括文本清洗、分词、英文单词标准化等技术[33-35]。然而，现有的针对中文文本的预处理方法并不适用于面向弹幕文本的颜文字检测与提取任务。例如，传统的文本清洗会将颜文字中的广义符号误剔除，传统的分词方法也会在颜文字的特殊符号的影响下进行错误分词，影响颜文字的提取效果。

鉴于此，针对弹幕文本清洗和分词，本文在 Emoly 中设计了针对颜文字检测与提取任务的文本预处理方法，主要包含文本清洗与候选字符串提取两个步骤。

#### 4.1.1　文本清洗：基于颜文字特征重定义的停用词去除

文本清洗旨在去除文本中的干扰噪声，获得干净的数据，包括停用词去除、全半角转换、大小写转换等方法[33]。本小节根据弹幕文本及颜文字特性进行文本清洗，并将其分为针对符号字符与针对非符号字符的清洗。

文本清洗的主要思路是基于颜文字特征重定义的停用词去除方法。停用词指在一段文本中高频率出现、却无实际意义或信息量很低的字符，例如英文的“the”“a”与中文的“的”“个”。在传统文本中，“，”“（”等符号字符大多作为分隔符或修饰符存在，具有高频率、低价值的特征，一般被划分至停用词类别中，并被大多数学术界流行的通用中文停用词表[1)]收录。考虑到广义符号在颜文字中的重要性，本文在传统停用词的基础上，结合字符的前后搭配字符等特征，重新对停用词进行定义，并以此进行针对符号字符与非符号字符的清洗。

对于非符号字符，将停用词定义为未曾在搜狗颜文字中出现过的字符。由于弹幕文本大部分都是由非符号字符构成，但其在颜文字组成字符中仅占约 10.5% 的比例，因此对于颜文字而言，大部分非符号字符属于“出现频率高但实际价值低”的停用词。同时，对于已在搜狗颜文字集中出现过的非符号字符，不将其作为停用词（如数字“3”、字母“T”等）。

对于符号字符，将停用词定义为用于分隔或修饰非符号字符停用词的字符，例如位于非符号字符停用词前后的“，”、成对的“（”“）”等符号。这些字符由于前后搭配非符号字符停用词，因此依旧行使了传统标点符号的分隔与修饰作用，可以作为停用词剔除；而若该符号字符一侧搭配有符号字符或属于非停用词的非符号字符，则它有概率与搭配字符共同构成颜文字，因此不能作为停用词剔除。

#### 4.1.2　候选字符串提取：基于 $N$-gram 的方法

颜文字是由多个字符构成的整体。基于文本分词方法的思想，首先将弹幕分割成多个可能包含颜文字的子串，再从中进一步提取颜文字。

鉴于组成颜文字的字符在物理位置上具有相互邻接的特征，$N$-gram 方法[36]可以在切分句子的同时保留字符间的邻接顺序，这恰巧适合含颜文字文本的分割。

因此，我们使用 $N$-gram 方法将清洗后的弹幕文本进行切分，以形成一个颜文字的初步候选字符串集合，并将其作为后续步骤的输入数据。$N$ 的数值根据颜文字长度特征选取，以确保将文本中的颜文字全部提取出来，避免 $N$ 值过小造成的颜文字遗漏与过大带来的额外检测量及误差。

1) https://github.com/goto456/stopwords

### 4.2 颜文字甄别与筛选:统计指标设计

#### 4.2.1 搭配固定程度指标:双向条件概率

颜文字内部子串之间存在一些“固定搭配”。当某个特殊子串出现时,大概率意味着对应特殊颜文字也会出现,且该概率会随子串长度增加而增大。以颜文字“(ˊ▽ˋ)”为例,绝大多数情况下,其子串“(ˊ▽ˋ”会与“)”搭配出现,而不会单独出现或与其他字符搭配出现,即字符串“(ˊ▽ˋ”出现时,颜文字“(ˊ▽ˋ)”有极大的概率会出现。基于这点,我们提出双向条件概率来描述在给定子串出现下,特定颜文字出现的概率。

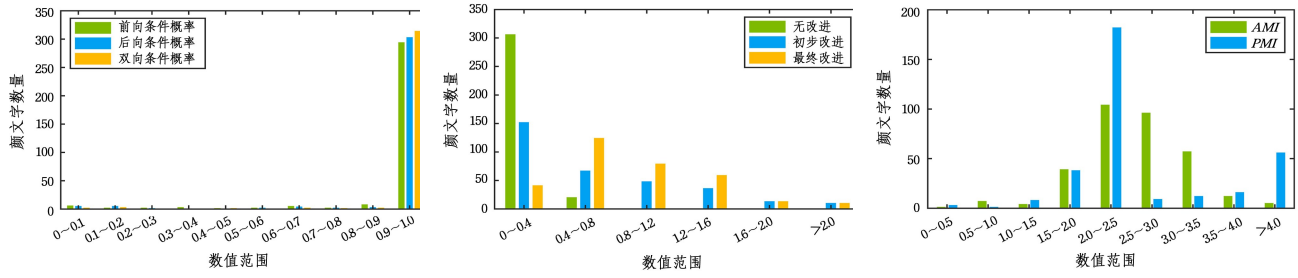
定义长度为 $n$ 的候选字符串 $S$ 的前向条件概率 $PR_f$ 与后向条件概率 $PR_b$ 的计算式分别如式(1)、式(2)所示,双向条件概率 $PR$ 的定义如式(3)所示。其中,$PR_f(S)$ 代表 $S$ 的前 $n-1$ 长度子串出现时,$S$ 出现的概率;$PR_b(S)$ 代表 $S$ 的后 $n-1$ 长度子串出现时,$S$ 出现的概率;双向条件概率 $PR(S)$ 代表 $S$ 的任意长度为 $n-1$ 的子串出现时,剩下的一个字符出现(也就是 $S$ 出现)的概率的最大值。候选字符串的双向条件概率越大,意味着该候选字符串是一个固定搭配组合的概率越大,亦即它属于颜文字的概率越大。

$$PR_f(S)=p(C_n|C_1C_2\cdots C_{n-1})=p(C_1C_2\cdots C_n)/p(C_1C_2\cdots C_{n-1}) \tag{1}$$

$$PR_b(S)=p(C_1|C_2C_3\cdots C_n)=p(C_1C_2\cdots C_n)/p(C_2C_3\cdots C_n) \tag{2}$$

$$PR(S)=\max(PR_f(S),PR_b(S)) \tag{3}$$

为了验证该指标的有效性,我们计算搜狗颜文字集中每个颜文字的前向、后向及双向条件概率并进行统计,如图2(a)所示。




(a)双向条件概率指标 (b)改进的左右信息熵指标 (c)AMI 与 PMI 指标

图 2 统计指标验证图

Fig. 2 Statistical indicators verification

可以看到,绝大多数颜文字的双向条件概率都位于0.9~1之间。另外,横向比较前向、后向与双向条件概率可以发现,颜文字的双向条件概率比单独的条件概率更高、更具有集中性与代表性。这一统计结果说明了双向条件概率用于颜文字检测的合理性与可行性。

#### 4.2.2 片段问题解决指标:左右信息熵

我们发现,双向条件概率也会筛选出许多非颜文字的字符串,人工检查这些错误项,发现它们可以分为两类。1)颜文字的不完整子串,即颜文字片段。例如“(ˊ▽ˋ”,它是颜文字“(ˊ▽ˋ)”的片段。通过观察发现,这类错误项往往占据较大比例。2)其他非颜文字的字符组合。这些字符组合往往由于频数过低或用户使用习惯,形成了固定搭配现象,拥有较高的双向条件概率,例如字符串“?!?!?”,但其并不属于颜文字。第二类错误项会在 4.2.3 小节阐述解决方法。本小节主要解决第一类错误项,即颜文字片段问题。出现这类问题的主要原因是除了颜文字(尤其是对于较长的颜文字)本身外,这些颜文字片段往往也具有较强的独特性,拥有较高的双向条件概率,导致被错误地提取出来,需要再次剔除。

为了解决这个问题,基于传统信息熵的概念,本文提出了改进的左右信息熵。初步改进的左右信息熵 $H_1$ 的计算式如式(4)所示,其中 $S+$ 与 $S-$ 分别代表字符串 $S$ 左侧与右侧的字符集合,不同的句首句末具有不同的字符表示;$p(x)$ 代表字符 $x$ 在清洗后的文本中出现的概率。该指标用于衡量候选字符串前后搭配字符的丰富程度,剔除只搭配固定字符的颜文字片段。左右信息熵越大,意味着该字符串具有越丰富的前后搭配词,有更大的概率是完整的颜文字;左右信息熵越小则意味着该字符串至少有一侧搭配字符单一,极有可能与其他字符共同构成一个固定搭配,是颜文字片段的概率越高。

$$H_1(S)=\min\left(-\sum_{x\in S^+}p(x)\log p(x),-\sum_{x\in S^-}p(x)\log p(x)\right) \tag{4}$$

$$H_2(S)=\min\left(-\sum_{m,n\in S^+}(p(m)\log p(m)+v\times p(n)\log p(n)),-\sum_{m,n\in S^-}(p(m)\log p(m)+v\times p(n)\log p(n))\right) \tag{5}$$

图 2(b)给出了对搜狗颜文字集中的颜文字应用未改进与初步改进的左右信息熵指标得到的结果。可以看到,尽管初步改进的左右信息熵计算结果相比未改进时有显著的提高,但整体数值依旧偏低,有 67% 的颜文字在 0.8 以下。对这些信息熵较低的颜文字进行排查,结果显示,在所有信息熵低于 1 的颜文字中,有 88.7% 的颜文字频数低于 10。大多数颜文字拥有着较低的频数,低频数导致了邻接符号相对贫乏,从而导致信息熵偏低。为了解决此问题,我们对信息熵计算方法再次进行结合频数与位置信息的改进,提出了一种最终改进的加权左右信息熵计算方法:当字符串频数大于设定的阈值时,依旧使用式(4)来计算;而当字符串频数小于该阈值时,则应用式(5)进行计算。式(5)中,$m$ 代表除首尾外的其他邻接字符,$n$ 代表首尾邻接字符,$v$ 代表为首尾字符增加的权重。当权重值设为 1 时,式(5)等价于式(4)。

式(5)的改进之处主要是放大了句首末字符串的信息熵增益程度。图 2(b)给出了对颜文字应用最终改进信息熵的计算结果,其中频数阈值设定为 10,权重设为 3。可以看到,最终改进的信息熵结果比先前有明显的提升。同时,我们

计算颜文字所有子串的最终改进的信息熵，发现有 85% 的子串的左右信息熵在 0.5 以下，平均值为 0.34。由此观察可得，绝大多数颜文字最终改进的信息熵都较高，而其片段的信息熵都偏低；最终改进的左右信息熵指标不仅能够更好地区分颜文字与颜文字片段，而且还能降低颜文字频数过低造成的误差。

#### 4.2.3　凝合程度指标：互信息

4.2.2 小节提到，双向条件概率还会将一些非颜文字的其他字符组合一起筛选出来，这是因为双向条件概率仅考虑字符串 $n-1$ 长度的子串与其他字符的搭配关系，而没有考虑字符串内部其他子串与字符之间的搭配关系与凝合程度；而这些非颜文字的字符组合，例如字符串“?!?!（”具有一个统一特点，即它们大多是由常见的符号字符随意拼凑而成，字符间的凝聚度极低。

考虑到这个特点，我们引入了平均互信息（Average Mutual Information，AMI）与点互信息（Pointwise Mutual Information，PMI）指标，用于量化字符串整体的凝合程度。其中，AMI 主要衡量字符串各个字符之间的凝合程度，而 PMI 主要衡量字符串各子串之间的凝合程度。

式(6)和式(7)分别为长度为 $n$ 的字符串 $S$ 的 AMI 与 PMI 的计算公式，其中分子代表 $S$ 在清洗后的文本中出现的概率，式(6)的分母代表组成 $S$ 的字符在清洗后的文本中出现概率的乘积，式(7)的分母则代表组成 $S$ 的子串在清洗后的文本中出现概率的乘积。这两个公式从字符和子串的出现概率这个角度，衡量了一个字符串内部字符或子串的凝合程度。

$$AMI(S)=\frac{1}{n}\log_2\frac{p(C_1C_2\cdots C_n)}{p(C_1)p(C_2)\cdots p(C_n)} \tag{6}$$

$$PMI(S)=\min\ \log_2\frac{p(C_1C_2\cdots C_n)}{p(C_1\cdots C_i)p(C_{i+1}\cdots C_n)},\quad 1\leqslant i\leqslant n \tag{7}$$

图 2(c)给出了搜狗颜文字集中所有颜文字的 AMI 与 PMI 统计结果。可以看到，大多数颜文字都具有较高的 AMI 与 PMI。同时，为了说明互信息指标对于检测非颜文字字符串的有效性，我们随机抽取部分通过前两个指标筛选出来的非颜文字字符串，计算其 AMI 与 PMI。表 2 列出了部分非颜文字字符串各类指标的结果。可以看到，前两项指标无法剔除的字符串大多有着较低的 AMI 与 PMI，可以通过互信息指标剔除。

表 2　不同字符串的各指标数值示例

Table 2　Example of indicators' values in different strings

| 字符串 | 双向条件概率 | 左右信息熵 | AMI | PMI |
|---|---|---|---|---|
| !!!】 | 1.0 | 1.079 181 246 047 624 7 | 0.281 264 046 131 25 | 0.463 983 985 337 428 |
| ?!?!（ | 1.0 | 0.903 089 986 991 943 4 | −0.102 516 054 056 09 | −0.733 172 155 821 87 |
| )…… | 0.88 | 1.477 121 254 719 662 4 | 0.218 541 903 859 370 | −1.431 084 871 534 40 |

### 4.3　基于颜文字惯用法的筛选：过滤规则设计

颜文字除了拥有上述统计特征外，在构成规则上也有一定的规律可循。基于对颜文字组成字符的分析和观察，我们提出了几点过滤规则，用于更好地提高算法效率，剔除错误项，降低误差。具体包含以下 3 条规则：

规则 1　剔除仅含同一个字符（空格除外）的字符串。

规则 2　剔除仅含一类非符号字符且不含符号字符（空格除外）的字符串。

规则 3　剔除仅含同一类非符号字符和标点类字符的字符串。

### 4.4　评估与排名指标：文本相似度

除了上述方法，利用候选字符串和已知颜文字间的相似度也能提取新颜文字。这一问题可以转化为短文本相似度问题。接下来我们将讨论两类文本相似度，并比较他们在颜文字检测与提取任务上的效果。

#### 4.4.1　基于文本字符的文本相似度：ROUGE 指标与雅卡尔相似系数

本文中基于文本字符的相似度计算主要包括 ROUGE 指标与雅卡尔相似系数两种方法。

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）通过词的共现信息来评价摘要[37]。本文选用 ROUGE-2 方法来比较候选字符串与基准颜文字子串之间的共现信息，计算式如式(8)所示。

$$ROUGE2\_score=\frac{|ROUGE(b)\cap ROUGE(c)|}{|ROUGE(b)|} \tag{8}$$

其中，$ROUGE(b)$ 与 $ROUGE(c)$ 分别代表基准颜文字与候选字符串通过 $N$-gram 方法获取的子串集合，$N$ 值取 2；分母代表基准颜文字的子串总数，分子代表候选字符串与基准颜文字相同子串的数量。通过计算候选字符串与所有基准颜文字的 ROUGE-2 分数并取其最大值，可以得到该候选字符串基于 ROUGE-2 方法的相似度分数。

雅卡尔相似系数（Jaccard Similarity Coefficient，简称 Jaccard）用于衡量两个样本集的相似程度，计算式如式(9)所示。

$$Jaccard\_score=\frac{|Jaccard(b)\cap Jaccard(c)|}{|Jaccard(b)\cup Jaccard(c)|} \tag{9}$$

其中，$Jaccard(b)$ 与 $Jaccard(c)$ 分别是候选字符串与基准颜文字所有组成字符构成的集合；分母与分子分别是两个字符集合的并集与交集。通过 Jaccard 方法计算候选字符串与所有基准颜文字的相似度分数并取其最大值，用于比较候选字符串与基准颜文字组成字符之间的重合程度。

#### 4.4.2　基于词向量的文本相似度：BOW 模型

基于词向量的相似度计算主要包括两部分，分别是词向量表示方法与相似度度量方法的选择。本文选择采用词袋模型（Bag-of-words Model，BOW）来表示词向量，并采用余弦相似度方法进行相似度度量。

通过 BOW 将基准颜文字与候选字符串向量化后，计算两者的余弦相似度，可以得到两者基于词向量的相似度，计算式如式(10)所示。

$$BOW\_score=\frac{vec(b)\cdot vec(c)}{|vec(b)||vec(c)|} \tag{10}$$

其中，$vec(b)$和$vec(c)$分别代表基准颜文字与候选字符串经BOW转换后的向量表示。通过计算候选字符串与所有基准颜文字基于词向量的相似度并取其最大值，可以得到该候选字符串最终基于BOW的相似度分数。

## 5 实验设计与评估

### 5.1 参数分析与设定

本小节首先介绍Emoly算法中所涉及的参数及其设定方法。表3列出了参数表示及其说明。其中，参数1适用于4.1.2小节的候选字符串提取阶段，将大于1、小于等于该参数的数值依次作为$N$-gram方法中的$N$值，对清洗后的文本进行分词，提取初步候选字符串。该参数的设定参考基准颜文字集中颜文字的长度特征。参数2是对候选字符串出现频数的最低要求。根据观察，大多数冷门颜文字在弹幕文本中的出现频数仅为1，因此本实验对该参数不作限制，即仅将其设置为1。参数3—参数8均在4.2小节的统计指标中被提及，是对候选字符串的进一步筛选；其数值通过计算搜狗颜文字在这些参数上对应的数值来确定，原则是确保所设定的参数数值能够有效区分颜文字与非颜文字的候选字符串，并且能够最大限度地将颜文字筛选出来。

表3 算法参数表示及说明

Table 3 Description of algorithm parameters

| 序号 | 参数名称 | 参数表示 | 说明及用途 |
|---|---|---|---|
| 1 | 候选字符串长度阈值 | sup(gram) | 候选字符串的长度上界，用于剔除过长与过短的字符串 |
| 2 | 频数阈值 | inf(count) | 候选字符串的频数下界，用于剔除频数过低的字符串 |
| 3 | 双向条件概率阈值 | inf(cond_PR) | 候选字符串的双向条件概率下界，用于剔除非固定搭配的字符串 |
| 4 | 左右信息熵阈值 | inf(entropy) | 候选字符串的左右信息熵下界，用于剔除颜文字片段 |
| 5 | 信息熵放大阈值 | inf(entropy_num) | 适用改进信息熵的频数下界，避免频数过低带来的误差 |
| 6 | 信息熵放大倍率 | v | 改进信息熵的放大倍率，避免频数过低带来的误差 |
| 7 | 平均互信息阈值 | inf(ami) | 候选字符串的平均互信息下界，用于剔除凝聚程度低的字符串 |
| 8 | 点间互信息阈值 | inf(pmi) | 候选字符串的点间互信息下界，用于剔除凝聚程度低的字符串 |

### 5.2 实验数据及评价基准

接下来，我们对本次实验所使用的数据集及评价基准进行详细介绍。

为了验证算法的有效性，我们在3.2小节所述的弹幕数据集中选取其中3个分区的弹幕数据，组成实验所用的弹幕数据集。所选择的3个分区分别为生活区、游戏区与鬼畜区。表4列出了对应数据集的介绍。

表4 弹幕数据集的介绍

Table 4 Description of danmaku datasets

| 弹幕数据集 | 弹幕数量/条 | 视频数量 | 时间范围 |
|---|---|---|---|
| 生活区(living) | 1 260 602 | 5 459 | 2018-10—2021-12 |
| 游戏区(gaming) | 1 910 230 | 6 361 | 2019-05—2021-12 |
| 鬼畜区(kichiku) | 485 868 | 3 098 | 2018-10—2021-12 |

生活区[1]视频大多是对日常生活的记录，其弹幕有着广泛的用户群体、丰富的弹幕资源，以及贴近日常交流的语言风格。

游戏区[2]视频大多围绕各种电脑手机游戏进行，其弹幕特点是有大量游戏领域相关的专有名词术语，用于研究算法在用户群体小众、领域性强的弹幕文本下的表现。本文选取了游戏区内一款日活跃用户数百万级别的手游视频专区弹幕作为实验数据。

鬼畜区[3]视频通常以恶搞为主，其弹幕内容创新性强，是许多网络热词的发源地。

同时，为了更好地验证Emoly算法检测与提取颜文字的效果，鉴于上文已使用搜狗颜文字集进行参数设置，我们采用百度输入法内置颜文字数据集(下文简称百度颜文字集)对算法召回率进行评估。百度输入法是百度有限公司推出的一款人工智能输入法工具[4]，截至2022年5月，在华为应用市场内拥有6亿下载量。我们收集了百度输入法内置的共693个颜文字，囊括开心、惊讶等多个类别。通过分析发现，百度颜文字集与搜狗颜文字集的交并比仅为5.27%，这不仅体现了颜文字的多样性，也说明了以百度颜文字集作为评价基准的科学性。

### 5.3 实验方法及结果评估

为了从多方面对Emoly算法进行评估，本文进行了一系列实验，用于回答以下研究问题。

RQ1：Emoly算法在弹幕数据集上是否具有准确性与有效性？

RQ2：Emoly算法能否在弹幕数据集上保持结果的稳健性？

RQ3：Emoly算法在发现新颜文字的能力上表现如何？

RQ4：与其他颜文字检测和提取算法相比，Emoly算法是否具有优越性？

#### 5.3.1 RQ1：Emoly算法在弹幕数据集上是否具有准确性与有效性？

本小节通过分别计算Emoly算法在生活区、游戏区、鬼畜区弹幕文本上的精确率(Precision，P)与召回率(Recall，R)，来验证其准确性与有效性。我们通过4.4小节的3种相似度计算方法分别对结果进行排序，并按照排序结果由高至低分段进行精确率与召回率的计算。由于目前并无收录全部颜文字的词库，仅依靠现有颜文字词库计算精确率难免会造成误差，因此精确率通过人工检测候选字符串是否属于颜文字来计算。考虑到数据集规模以及人工检测的局限性，同时考虑现有工作[19]采用的评估方法，我们仅手动评估排名前1 000的候选字符串。召回率则是以百度颜文字集为基准，计算在存在于弹幕文本的所有百度颜文字中，被算法提取出来的颜文字所占的比例。表5列出了生活区的实验结果(游戏

[1] https://www.bilibili.com/v/life

[2] https://www.bilibili.com/v/game

[3] https://www.bilibili.com/v/kichiku

[4] https://shurufa.baidu.com

区和鬼畜区的结果类似，已公开在 figshare[21] 上），记录了在 3 种相似度计算方法下，以 100 为间隔数值时，排名在此之前的候选字符串的精确率、召回率及该名次候选字符串对应的相似度分数。

在 3 个数据集中，排名前 1 000 的候选字符串均能达到 96%以上的精确率。其中，最高的精确率均通过 Jaccard 方法来达到。在 3 种相似度算法中，Jaccard 方法取得了最好的结果。对于产生这一结果的原因，推测如下：1）Jaccard 方法仅考虑颜文字与候选字符串字符的重合度，ROUGE-2 方法将字符顺序也加入了考量，而 BOW 方法使用全部候选字符串与基准颜文字字符进行向量建模；2）Jaccard 排名方法考虑到数据集文本规模、颜文字的丰富度与其自身构成的独特性，更加具有优越性与准确性。当 Jaccard 相似度分数在 0.6 及以上时，精确度可以达到 99%及以上，这表明候选字符串的相似度分数越高，其属于颜文字的概率就越大；而当 Jaccard 相似度仅为 0.4 左右时，依然能够达到 96%左右的精确度。这一方面说明了相似度指标的有效性，可以通过相似度分数更加方便地对字符串属于颜文字的概率进行量化，同时也说明了算法的有效性，能够准确提取出具有新结构、与已知颜文字相似度低的新型颜文字。因此，在后续实际应用中，可以将相似度阈值初步设定为 0.4，并根据排名靠后的字符串是否属于颜文字来进行调整。

对结果的召回率进行评估，可以看到，当以百度颜文字为评价基准时，Emoly 算法在 3 个数据集上的召回率都能达到 91%及以上，这表明 Emoly 算法能够有效且较为全面地将弹幕文本中存在的颜文字提取出来。

表 5　Emoly 在生活区语料库上的精确率和召回率的实验结果

Table 5　Experimental results of Emoly's precision and recall on the living corpus

| | | | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 生活区 living corpus | score | ROUGE-2 | 1.00 | 1.00 | 1.00 | 0.77 | 0.60 | 0.50 | 0.43 | 0.36 | 0.29 | 0.22 |
| | | BOW | 1.00 | 1.00 | 1.00 | 0.91 | 0.86 | 0.84 | 0.80 | 0.77 | 0.75 | 0.70 |
| | | Jaccard | 1.00 | 1.00 | 1.00 | 0.80 | 0.67 | 0.60 | 0.57 | 0.50 | 0.50 | 0.43 |
| | *P*/% | ROUGE-2 | 100 | 100 | 100 | 99.5 | 99.0 | 97.8 | 97.8 | 97.5 | 96.9 | 95.5 |
| | | BOW | 100 | 100 | 100 | 99.5 | 98.2 | 95.7 | 94.7 | 93.3 | 91.1 | 89.3 |
| | | Jaccard | 100 | 100 | 100 | 100 | 99.6 | 99.3 | 99.0 | 98.3 | 97.6 | **97.0** |
| | *R*/% | ROUGE-2 | 30 | 60 | 90 | **93** | 93 | 93 | 93 | 93 | 93 | 93 |
| | | BOW | 32 | 61 | 88 | **93** | 93 | 93 | 93 | 93 | 93 | 93 |
| | | Jaccard | 32 | 61 | 87 | **93** | 93 | 93 | 93 | 93 | 93 | 93 |

相比之前基于词典匹配、机器学习与视觉特征方法的其他颜文字检测与提取工作，Emoly 算法在准确率与召回率上依旧具有优势，我们会在之后的实验中将 Emoly 算法与其他工作进行对比来说明这一点。Emoly 算法发现的所有 4 732 个颜文字已开源[21]。

### 5.3.2　RQ2：Emoly 算法能否在弹幕数据集上保持结果的稳健性？

为了验证 Emoly 算法在不同数据集上的效果是否具有稳健性，本小节通过将算法应用到不同数据集并比较提取出来的结果的重合度，来说明这一点。首先，我们使用 5.2 小节所述弹幕数据集，构造了在分区类别、数据内容、数据规模上各有差异的 12 个小规模数据集，构造步骤如下：

步骤 1　从 5.2 小节中的 3 个分区的数据集中各随机取出 25 万条弹幕数据组成 3 个弹幕数据集，并将其简单命名为游戏区 I、生活区 I 与鬼畜区 I 数据集。

步骤 2　对于弹幕规模为 25 万的游戏区 I 数据集，我们从中随机取出 10 万条数据并将此过程重复 3 次，得到 3 个规模为 10 万的数据集，将其分别命名为游戏区 II、游戏区 III 与游戏区 IV 数据集。

步骤 3　将生活区 I 与鬼畜区 I 数据集分别执行步骤 2，得到规模为 10 万的生活区 II、生活区 III、生活区 IV 与鬼畜区 II、鬼畜区 III、鬼畜区 IV 数据集。

将 Emoly 算法分别应用到上述 12 个数据集中，并对得到的结果按照不同分区与规模进行结果重合度的计算，重合度即算法在不同数据集输出的结果中相同字符串占结果总数的比例。将 3 个不同分区的数据集进行两两对比。图 3 给出了不同分区数据集下的结果重合度，横坐标代表按照相似度分数得出的排名，纵坐标代表在两个不同数据集中，排名在对应横坐标之前的算法结果的重合度，并按照名次变化绘制重合率曲线。这里使用 ROUGE-2 计算相似度分数。

可以看到，随着名次的逐步下降，所有重合度曲线均呈下降趋势，且相同分区内的重合度曲线几乎保持一致。当名次在 70～100 区间内时，结果重合度均最高，可以达到 0.7 及以上；而对于同属于游戏区与生活区的数据集而言，重合度甚至可以达到 0.9。这意味着，算法从不同数据集中提取出的排名前列的结果大部分是相同的。对于属于同一分区的不同数据集，最终结果重合度曲线的变化趋势也类似，对于其他数据集的分析，我们放在 figshare 上作为补充材料[21]。上述实验表明了 Emoly 算法可以稳健地从数据集中提取出颜文字，当同一个颜文字同时存在于不同数据集中时，算法均可以稳定地将其提取出来，而不会受到具体数据内容的干扰。

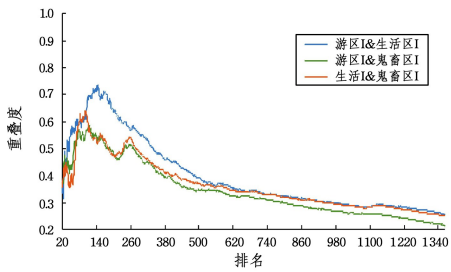



图 3　25 万规模三分区弹幕数据集上结果重合度随排名变化的曲线

Fig. 3　Curve of result overlap with ranking changes on a two hundred and fifty thousand-scale three-partition danmaku dataset

5.3.3 RQ3:Emoly 算法发现新颜文字的能力如何?

颜文字检测与提取算法的一大应用就是发现新颜文字,从而指导输入法词库更新以及对颜文字发展趋势的研究等。

本小节以 5.3.1 小节的实验为基础,以百度颜文字库为基准,将 Emoly 算法发现的且未被百度颜文字库收录的颜文字定义为新颜文字。

统计结果发现,在 Emoly 检测和提取到的 4 869 个颜文字中,仅有 419 个是被收录的,而剩下的 4 450 个颜文字都属于新发现的新颜文字,占总发现颜文字的 91.39%。可以看出,Emoly 在发现新颜文字上的效果较好。

考虑到百度颜文字库仅包含 693 个颜文字,数量较少,为了尽可能减少现有输入法颜文字库数据量带来的影响,我们进一步扩大了基准颜文字库。我们收集了在华为应用市场中下载量均在亿级的 4 个手机输入法(搜狗输入法、百度输入法、讯飞输入法、QQ 输入法)中的词库,他们分别包含 1 144 个、693 个、1 573 个、983 个颜文字(合并去重后共 3 854 个)。经统计发现,在 Emoly 发现的 4 869 个颜文字中,也仅有 773 个被收录在这个包含 4 450 个基准颜文字的库中,换言之,新发现的颜文字占总发现颜文字的 84.12%。

以上分析表明,Emoly 在新颜文字的发现上表现较好。接下来,我们以 1 个 Emoly 发现的新颜文字案例来探究 Emoly 算法在新颜文字发现上有这样表现的原因,发现在 Emoly 中包含罕见 Unicode 字符的颜文字。在 Emoly 发现但未被收录的颜文字中,有一些包含罕见 Unicode 字符的颜文字。例如"ɞ♡ɞ"是一个带有翅膀的心,其中组成翅膀的"ɞ"和"ɞ"是拉丁文字母,Unicode 编号分别为 U+029A 和 U+025E,而"♡"是一个 emoji 符号,Unicode 编号为 U+2764。再比如,"ϙ⍢ϙ"是一个举起双拳的笑脸表情,其中组成双拳的"ϙ"和"ϙ"是奥里亚文数字,Unicode 编号分别为 U+0B67 和 U+0B68,而"⍢"是 APL 编程语言中的一种函数符号,Unicode 编号为 U+2362。由于地域、文化等原因,这些罕见的 Unicode 字符在日常生活中用得极少,却在颜文字中大放异彩,重新被赋予了新的含义。

从上述案例可以看出,由于地域文化等原因,输入法无法收录到部分含有罕见字符的颜文字,加之颜文字可以灵活添加组件(见 5.3.3 小节中提到的 AZEmo[16]),以及创作者源源不断地创造新颜文字,输入法难以收录含义完全或结构相似的颜文字。

5.3.4 RQ4:与其他颜文字检测与提取算法相比,Emoly 算法是否具有优越性?

本小节通过将 Emoly 算法与 Yu 等提出的 AZEmo 系统[16]进行对比,来体现 Emoly 算法的优越性。AZEmo 系统的核心思想是利用人体运动学理论,将已知颜文字分割为多个组件,并利用组件词典对文本进行匹配来提取颜文字。基于上述思想与文献中描述的具体步骤,我们对 AZEmo 系统进行复现:首先使用 3.1 小节中的搜狗颜文字集对组件词典进行构造;然后使用构造的组件词典依次对弹幕文本进行匹配,并使用正则方法提取匹配到的字符串,并将其作为结果输出。我们依然使用百度颜文字集作为召回率评价基准,使用生活区数据集对 AZEmo 系统与 Emoly 算法所得结果的精确率与召回率进行计算,相似度采取 Jaccard 方法。表 6 列出了两种算法的对比结果。

表 6 两种算法在生活语料库上的颜文字检测结果

Table 6 Results of kaomoji detection of two algorithms on living corpus

(%)

| methods | *Precision* | *Recall* |
|---|---|---|
| Emoly | 97(top 1 000) | 93 |
| AZEmo with Sogou dict | 65.2 | 60 |
| AZEmo with Emoly's output dict | 54.2 | 95 |

当使用搜狗颜文字集构建词典时,AZEmo 系统在生活区弹幕文本中颜文字的召回率仅为 60%,有较多的颜文字未被检测出来。相比之下,Emoly 算法具有明显的优势。对 AZEmo 系统的结果进行分析发现,其误差主要源于两方面:首先,该算法的效果强烈依赖于先前构造的组件词典,难以提取出组件词典中不存在的、拥有全新组成字符与结构的颜文字。其次,正则提取时将颜文字以外的字符一起作为结果提取出来,也是造成误差的原因之一。Emoly 算法通过结合统计指标与过滤规则,大大减弱了对已有颜文字的依赖性,从而可以较为全面地将颜文字从文本中提取出来,拥有较高的召回率。同时,我们使用 Emoly 算法检测出的排名前 1 000 的颜文字再次构造组件词典,重新计算 AZEmo 系统的精确率与召回率。在应用 Emoly 算法提取出的颜文字作为词典后,AZEmo 系统的召回率得到了显著提升;然而,其精确率也从 65.2%下降到了 54.2%。这一结果进一步说明了 Emoly 算法的优越性与应用价值。

除了与上述基于词典的颜文字提取方法进行对比外,表 7 还列出了 Emoly 算法分别与基于机器学习以及基于视觉特征的相关工作在精确率与召回率上的对比结果,表 7 中的数值均来源于对应文献。Tanaka 等[29]借鉴了自然语言处理中的组块分析(Chunking),首先结合形态学信息,对每个文本中的每个字符做类型标记,然后将颜文字的提取任务视为组块分析任务,用于提取文本中的颜文字。Takeru 等[31]关注到颜文字中"眼睛"这个重要视觉特征,提出了一套基于识别"眼睛"字符和字符串对称性的方法,首先检测出文本中表示颜文字"眼睛"的字符,然后比较左眼和右眼部分文本字符串的相似程度,从而判断和提取颜文字。Bedrick 等[19]从字形的角度入手,结合程序语言设计的思想,提出了一套关注符号字形特征和对称性特征的概率上下文无关文法(Probabilistic Context-free Grammar,PCFG)来提取文本序列中的颜文字。本文的实验结果表明,与这些先前的主流方法相比,本文提出的 Emoly 算法能够达到较高的精确率与召回率,具有优越性。

表 7 Emoly 算法与其他工作的对比结果

Table7 Comparison between Emoly algorithm and other algorithms

(%)

| Methods | *Precision* | *Recall* |
|---|---|---|
| Emoly(living corpus) | 97(top 1 000) | 93 |
| Tanaka 等[29] | 85.5 | 86.7 |
| Takeru 等[31] | 58.7 | 58.8 |
| Bedrick 等[19] | 94.5(top 1 000) | — |

## 6　应用场景展望及工程化实践

### 6.1　中文分词

Emoly 可以为中文分词工具提供颜文字支持。分词是自然语言处理的重要步骤，近年来，许多研究人员致力于中文分词任务的研究并取得了显著的成果[38-41]，开发出了许多成熟的中文分词工具包，如 jieba[1)]。这些工具包通常具有良好的分词效果，被广泛应用于中文文本处理工作[42-43]，但都没有考虑颜文字对分词的影响。

### 6.2　文本情感分析

目前，已经有许多研究人员关注到颜文字在文本情感分析中的重要作用，开始对颜文字的情感信息进行研究[16,24]，并进一步结合颜文字对文本情感进行分析[26]。然而，这些结合颜文字的情感分析工作大多直接使用来源于网络或是自己收集的颜文字词库，这些词库往往具有数量少、不够全面等局限性，无法将文本中的颜文字完全覆盖。Emoly 算法可以由此对情感分析工作进行优化。通过 Emoly 算法，可以较为全面地将文本中存在的颜文字检测出来，减少因词库的局限性而产生的缺漏；再结合相关颜文字情感信息检测方法与传统文本情感分析方法，进而形成一套完整有效的针对网络文本的情感分析方法，改善文本情感的分析效果。

### 6.3　输入法词库更新

目前，各大输入法软件都收录了自己整理和用户创造的颜文字库，但都存在更新速度慢、收录不全的特点。以搜狗输入法与百度输入法为例，两者分别收录了 1144 与 693 个颜文字，其中仅有 92 个相同的颜文字。不完整的颜文字词库会对用户的输入体验造成不良影响，使用户无法自由地调用所有网络中的颜文字，对该输入法造成负面影响。如何全面收录网络中的颜文字并及时更新颜文字词库，是对各大输入法软件的重要挑战。目前，Emoly 提供的颜文字词典可以为各大输入法的颜文字库提供支持[21]，Emoly 算法也能在未来不断发掘新颜文字，更新和丰富各大输入法的颜文字库。

**结束语**　随着颜文字在网络文本中愈发广泛的使用，其在文本中具有的巨大语义情感价值与现有颜文字提取技术的局限性，表明了迫切需要一种新的算法，能够有效、准确并全面地将颜文字从大规模的网络文本中提取出来。本文提出了一种基于大规模弹幕文本的颜文字检测与提取算法 Emoly，通过将针对弹幕文本特点的预处理操作、多种改进的统计指标与过滤规则以及相似度排名评估方法相结合，使其能够有效地从大规模弹幕文本中自动化提取颜文字，并通过实验验证了该算法的准确性、稳健性、优越性与通用性。最后对算法的一些细节、误差产生的原因与改进方向进行了探讨，并对算法在多个领域的后续应用场景进行了展望。

## 参 考 文 献


[1] China Internet Network Information Center. Statistical Report on Internet Development in China [EB/OL]. (2017-08-03) [2023-03-24]. https://cnnic.cn/n4/2022/0401/c88-1129.html.

[2] Wikipedia contributors. Danmaku [EB/OL]. (2023-02-24)[2023-03-24]. https://en.wikipedia.org/wiki/Danmaku.

[3] XIAN Y K, LI J F, ZHANG C X, et al. Video highlight shot extraction with time-sync comment [C]// Proceedings of the 7th International Workshop on Hot Topics in Planet-scale Mobile Computing and Online Social Networking. ACM, 2015: 31-36.

[4] XU L L, ZHANG C. Bridging video content and comments: Synchronized video description with temporal summarization of crowdsourced time-sync comments [C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, 2017: 1611-1617.

[5] WU B, ZHONG E H, TAN B, et al. Crowdsourced time-sync video tagging using temporal and personalized topic modeling [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 721-730.

[6] HE M, GE Y, WU L, et al. Predicting the popularity of danmu-enabled videos: A multi-factor view [C]// Proceedings of the 21st International Conference on Database Systems for Advanced Applications. Springer-Verlag, 2016: 351-366.

[7] WU F M, LV G Y, LIU Q, et al. Deep Semantic Representation of Time-Sync Comments for Videos [J]. Journal of Computer Research and Development, 2019, 56(2): 293-305.

[8] LV G Y, XU T, CHEN E H, et al. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding [C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, 2016: 3000-3006.

[9] HE M. Mining techniques for online videos' danmu data [D]. Hefei: University of Science and Technology of China, 2018.

[10] Wikipedia contributors. Emoticon [EB/OL]. (2023-04-20)[2023-05-15]. https://en.wikipedia.org/wiki/Emoticon.

[11] JING M. Kaomoji: Emojis and cultural representations in the Age of Reading Pictures [J]. Journal of Southwest University for Nationalities (Humanities and Social Science), 2020, 41(11): 149-155.

[12] DAANTJE D, ARJAN E, JASPER G. Emoticons and social interaction on the Internet: the importance of social context [J]. Computers in human behavior, 2007, 23(1): 842-849.

[13] JARAM P, VLADIMIR B, CLAY F, et al. Emoticon style: Interpreting differences in emoticons across cultures [C]// Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. AAAI Press, 2013: 466-475.

[14] CAO Z J, YE J. Attention savings and emoticons usage in BBS [C]// Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology. IEEE Computer Society, 2009: 416-419.

[15] Wikipedia contributors. Stop word [EB/OL]. (2023-03-13) [2023-03-24]. https://en.wikipedia.org/wiki/Stop_word.

[16] YU S, ZHU H Y, JIANG S, et al. Emoticon analysis for Chinese social media and e-commerce: The AZEmo system [J]. ACM Transactions on Management Information Systems, 2019, 9(4): 1-22.

[17] HOGENBOOM A, BAL D, FRASINCAR F, et al. Exploiting emoticons in polarity classification of text [J]. Journal of Web Engineering, 2015, 14(1/2): 22-40.


1) https://github.com/fxsjy/jieba

[18] YAMADA T, TSUCHIYA S, KUROIWA S, et al. Classification of facemarks using n-gram [C]// Proceedings of 2007 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2007: 322-327.
[19] BEDRICK S, BECKLEY R, ROARK B, et al. Robust kaomoji detection in Twitter [C]// Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics, 2012: 56-64.
[20] ZHAO X F, JIN Z G. Multi-dimensional sentiment classification of microblog based on Emoticons and short texts [J]. Journal of Harbin Institute of Technology, 2020, 52(5): 113-120.
[21] MAO X, LEI Z Y, XIA M Y, et al. The Emoticons Discovered by Emoly [EB/OL]. (2023-04-17) [2023-04-17]. https://figshare.com/articles/dataset/The_Emoticons_Discovered_by_Emoly/22639207.
[22] Wikipedia contributors. Emoticon [EB/OL]. (2023-03-07) [2023-03-27]. https://en.wikipedia.org/wiki/Emoticon.
[23] SONG Z X. Non-verbal Communication [M]. Shanghai: Fudan University Press, 2008: 1-18.
[24] PTASZYNSKI M, MACIEJEWSKI J, DYBALA P, et al. CAO: A fully automatic emoticon analysis system based on theory of kinesics [J]. IEEE Transactions on Affective Computing, 2010, 1(1): 46-59.
[25] CHEN X, ZHANG Y X, WU J C, et al. Construction and Analysis of Diachronic Bullet-screen Comment Corpus: Case Study of Youth Subculture Bullet-screen Comment [J]. Information Research, 2022, 2022(9): 77-85.
[26] LI Z, LI R, JIN G H. Sentiment analysis of danmaku videos based on Naïve Bayes and sentiment dictionary [J]. IEEE Access, 2020: 75073-75084.
[27] AHMAD S, VARMA R. Information extraction from text messages using data mining techniques [J]. Malaya Journal of Matematik, 2018, 5(1): 26-29.
[28] LIU L J. Research on text sentiment analysis for bullet screen [D]. Lanzhou: Lanzhou Jiaotong University, 2020.
[29] TANAKA Y, TAKAMURA H, OKUMURA M. Extraction and classification of facemarks [C]// Proceedings of the 10th International Conference on Intelligent User Interfaces. ACM, 2005: 28-34.
[30] KWON J, KOBAYASHI N, KAMIGAITO H, et al. Bridging between emojis and kaomojis by learning their representations from linguistic and visual information [C]// Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. ACM, 2019: 116-123.
[31] YOKOI T, KOBAYASHI M, IBRAHIM R. Emoticon extraction method based on eye characters and symmetric string [C]// Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2015: 2979-2984.
[32] Wikipedia contributors. Sogou Pinyin [EB/OL]. (2022-12-24) [2023-03-27]. https://en.wikipedia.org/wiki/Sogou_Pinyin.
[33] ALASADI S, BHAYA W. Review of data preprocessing techniques in data mining [J]. Journal of Engineering and Applied Sciences, 2017, 12(16): 4102-4107.
[34] LOSARWAR V, JOSHI M. Data preprocessing in web usage mining [C]// Proceedings of the International Conference on Artificial Intelligence and Embedded Systems. 2012: 15-16.
[35] LIU M J, WANG X F, HUANG Y L. Data preprocessing in data mining [J]. Computer Science, 2000, 27(4): 54-57.
[36] Wikipedia contributors. N-gram [EB/OL]. (2023-03-10) [2023-03-27]. https://en.wikipedia.org/wiki/N-gram.
[37] LIN C Y. ROUGE: A Package for Automatic Evaluation of Summaries [C]// Text Summarization Branches Out. 2004: 74-81.
[38] HUANG C N, ZHAO H. Chinese Word Segmentation: A Decade Review [J]. Journal of Chinese Information Processing, 2007, 21(3): 8-19.
[39] LUO R X, XU J J, ZHANG Y, et al. Pkuseg: A toolkit for multi-domain chinese word segmentation [EB/OL]. (2019-06-27) [2023-03-27]. https://doi.org/10.48550/arXiv.1906.11455.
[40] SUN M S, CHEN X X, ZHANG K X, et al. THULAC: An Efficient Lexical Analyzer for Chinese [EB/OL]. (2018-07-27) [2023-03-27]. https://github.com/thunlp/THULAC.
[41] SONG Y, CAI D F, ZHANG G P, et al. Approach to Chinese Word Segmentation Based on Character-Word Joint Decoding [J]. Journal of Software, 2009, 20(9): 2366-2375.
[42] YIN R C, WANG Q, LI P, et al. Multi-granularity chinese word embedding [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2016: 981-986.
[43] DAY M Y, LEE C C. Deep learning for financial sentiment analysis on finance news providers [C]// Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2016: 1127-1134.

**MAO Xin**, born in 1999, postgraduate. Her main research interests include natural language processing and data analysis.

**QI Zhengwei**, born in 1976, Ph.D, professor, Ph.D supervisor, is a member of CCF(No. 10710D). His main research interests include program analysis, model checking, virtual machines, and distributed systems.

(责任编辑:何杨)