基于 BERTopic 模型的 B 站学习视频弹幕主题挖掘和演化分析

Topic Mining and Evolutionary Analysis of Bilibili Learning Video Danmu Based on BERTopic Model

穆亚侨[1]，吴骏杰[1*]，王涛[2]，鲜隽桦[1]
[1]澳门理工大学应用科学学院
[2]华中师范大学人工智能教育学部
[*] junjiewu4-c@my.cityu.edu.hk

【摘要】 在线学习环境中，弹幕将学习者的实时讨论与视频内容紧密结合，为分析学习者学习状态提供了丰富数据资源。然而，现有弹幕分析方法主要依赖传统机器学习框架，未能充分挖掘弹幕文本语义特征，也未有效利用弹幕时序信息。本文基于 BERTopic 模型，对 B 站学习视频中的弹幕进行主题挖掘，分析其主题分布、讨论方向及演化趋势，旨在深入理解弹幕交互的整体表现。研究结果表明，弹幕不仅有助于学习者构建知识和提供情感支持，且能够反映学习者学习需求变化。这为优化在线学习资源的使用、提升教师教学设计及改进教学质量提供了有价值的参考。
【关键词】 在线讨论；弹幕；主题挖掘；BERTopic

Abstract: *In online learning environments, bullet comments (danmu) integrate real-time user discussions with video content, offering valuable data for analyzing learners' states. However, existing analysis methods rely on traditional machine learning frameworks, neglecting semantic features and temporal information. This paper uses the BERTopic model to mine topics from bullet comments in Bilibili learning videos, exploring their distribution, discussion direction, and evolution. Results show that bullet comments aid knowledge construction, emotional support, and reflect learners' needs and emotional shifts. These insights provide valuable guidance for optimizing online learning resources, enhancing teaching design, and improving teaching quality.*
Keywords: online discussion, danmu, topic mining, BERTopic

1. 引言

在线教育作为远程教育的实际应用，其本质上是教与学的时空分离，这一特点使得教与学的有效整合成为在线教育成功的关键，而在线交互正是实现这一整合的核心。在线交互是在线学习背景下学习群体之间的状态互动，尽管目前在线教育中通过传统的评论系统、学习社区等方式促进师生之间的交流，但这种交互形式约束条件较多，难以满足学生个性化的学习需求。

Bilibili 网站（B 站）的兴起，把“弹幕”这种允许演化实时分布和查看评论的新型交互模式在国内进行了推广，并迅速成为在线教学视频中用户进行实时交互行为的重要途径。Lin（2018）等人的研究表明弹幕独特的短文本和即时性特点使得用户可以使用简短的文本表达观点与情感，有效缓解在线学习者缺少学习互动和体验性不足等问题。

尽管在线学习的弹幕数据资源非常丰富，但弹幕评论内容短、口语化现象突出、网络流行用语较多，给针对弹幕的主题分析带来挑战。基于深度学习的 BERTopic 主题建模方法可以结合句子深层语义，为弹幕主题研究提供新的更优方法（王歌，2024）。因此本文选择采用 BERTopic 模型挖掘 B 站平台学习视频中弹幕主题交互现状、观点倾向与演化趋势，分析学习者的学习行为和状态变化。

2. 文献综述

2.1. 弹幕交互研究

在线教育由于其固有的特点，师生之间的互动提升面临着诸多挑战。穆肃等人（2020）等人根据等效交互原理，指出强化生生交互对优化在线学习体验、提高教育质量方面，具有重要的实践意义。弹幕交互作为学习者之间在线交互的主要沟通方式之一，已经开始被部分研究人员关注。张婧婧等人（2017）等人通过分析 B 站 PS 课程弹幕数据，发现学习者基于弹幕实现了语言认同，增强了学习群体的社会认知感，减轻孤独感。易田（2018）认为弹幕对于促进程序性知识学习起到积极作用。陈忆金等人（2021）指出与视频强相关的弹幕会提高学习者的认知存在感。张文兰等人（2022）的研究证明弹幕中的经验分享和知识总结等内容有效扮演了学习助手的角色，缓解学生学习焦虑。周璐（2024）认为弹幕是时空异步的生生互动工具，从社会存成方面为学习者跨时空提供情感支持。

现有研究主要聚焦于弹幕的情感特征分析，对弹幕在知识构建和认知生成方面的研究仍显不足。以在线编程课程为例，课程和弹幕中大量使用计算机专业英语单词，相关研究却忽视对弹幕英文主题特征的深入挖掘。因此，本文提出对英文文本进行单独的主题建模，以便更全面地探讨弹幕交互中的知识建构过程，深化在线学习过程中弹幕多维功能的理解。

2.2. 主题挖掘

主题挖掘源于对大规模文本数据中隐藏信息的需求，旨在从大规模文本数据中发现主题或话题的模式和结构。主题挖掘可以帮助研究人员分析和理解文本数据的主题发布情况、主题内容的关联性和学习群体的情感倾向等内容。常用的主题挖掘方法 LDA（Latent Dirichlet Allocation），即潜在狄利克雷分配于 2003 年由 Blei 等人（2003）提出。刘三女牙等人（2017）通过 LDA 模型自动挖掘和解析文本评论信息的特征结构和语义内容，为在线课程评论研究提供了思路。田园（2021）运用数据挖掘理论和 LDA 主题识别模型对在线教学需求数据进行主题挖掘，并构建多元化的在线教学用户需求指标体系。赖显静（2023）使用 LDA 主题模型对 MOOC 的评论文本进行分析，挖掘学习者的关注热点。王冲等人（2024）利用 LDA 模型结合情感分析，对 ChatGPT 类生成式人工智能在教育应用中的情感态度、关注热点与看法意见进行剖析。

尽管 LDA 模型在相关研究中被广泛应用，但其难以捕捉了词语语义和上下文关系，导致在处理短文本（如弹幕）时，存在主题内部联系不足和词汇代表性弱的问题（Egger & Yu, 2022）。此外，LDA 模型无法动态利用弹幕的时序特征，限制了对主题演化趋势的分析。相比之下，文本嵌入（text embedding）的出现(Devlin et al., 2018)，在生成文本词向量和特征表示方面取得了很好的效果，这种基于预训练模型生成的多维向量，可以从语义属性对文本的进行编码。基于深度学习的 BERTopic（Grootendorst, 2022）能够结合文本嵌入向量，提供更优的主题挖掘方法。已有研究（杨思洛与于永浩，2024）证明 BERTopic 在特定领域的主题挖掘上具有良好的分析效果。因此，本文选用 BERTopic 模型进行弹幕文本的主题识别，利用其强大的语义解释能力，更精准挖掘主题内在含义。

3. 研究过程

3.1. 数据介绍

本文选择 B 站人工智能类热门学习视频：吴恩达教授的“深度学习”公开课为研究对象。该课程被分为 183 个视频，截止 2024 年 11 月 27 日，所有视频总播放量达到 342 万，弹幕数量达到 3.1 万条。本文在数据采集阶段使用网络爬虫技术获取课程的弹幕文本，弹幕出现时间，弹幕发送 ID 等关键信息。通过数据清洗，删除重复、冗余、无意义等低质量数据后，最终保留 25254 条中文弹幕并从中筛选出 8359 条英文弹幕。

3.2. 弹幕主题挖掘

弹幕主题挖掘过程如图 1 所示，首先利用 jieba 分词工具对弹幕文本进行中文分词处理。通过该工具将弹幕中的文本切分为一个个具有意义的词语。在此过程中，还结合了中文文本分析工具中的停用词表，去除诸如“的”、“了”、“是”等常见的停用词，以提高文本分析的准确性。接着采用 BERTopic 模型进行主题挖掘。该模型首先对弹幕文本进行嵌入和降维计算，旨在将文本转化为数值化的向量表示，以便于后续的主题识别与分析。与传统主题建模方法不同，BERTopic 通过将同一主题下所有文档合并为一个长文档，基于整个主题集合的词频生成词袋表示，从而识别出该主题中的核心特征词。在主题挖掘过程中，模型通过计算主题之间的距离和余弦相似度，使用层次聚类将相似的主题聚集在一起，从而归纳出弹幕的核心讨论方向。此外，BERTopic 支持动态主题建模（Dynamic Topic Modeling，简称 DTM），能够结合时间戳信息，分析各主题随时间变化的演化趋势。

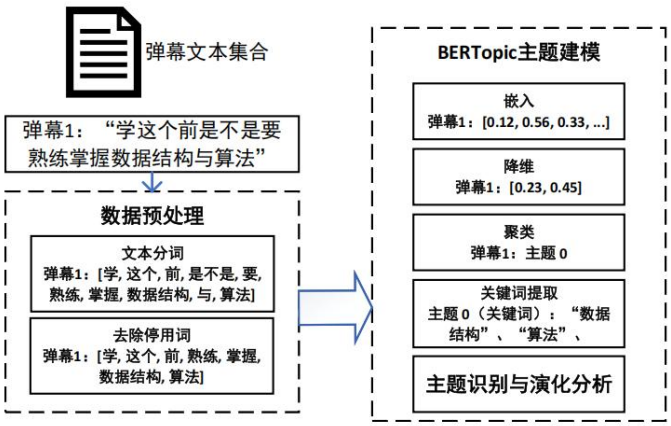

图 1 *弹幕文本主题挖掘*

4. 数据分析

利用 BERTopic 模型对 B 站视频的弹幕文本进行主题建模时，因为课程为英文授课且涉及到专业英文词汇，本文也将英文文本单独进行主题挖掘。根据多次实验结果调整模型参数，最终在表 1 列出学习者主要讨论的 20 个中文主题和 20 个英文主题。主题编号越小说明主题相关弹幕数量越多，讨论度越高。识别出的每个主题都由一组特征词来表示，这些词以不同的权重来代表主题。中文弹幕 Topic0 主题的文本占比达到 66.6%，其特征词是视频开始和结束时学习者互相问好和签到用语，与教学内容相关性较弱。英文主题中 Topic0 占比达到 90%，其特征词内容大多数学公式中各种英文简写。除 Topic0 外，中英文其余主题尽管占比较低，但都与课程内容相关性较强且具有代表性，证明了 BERTopic 模型的有效性。

表 1

主题识别结果和相关弹幕占比

中文主题&特征词	弹幕占比	英文主题&特征词	弹幕占比
T0（老师/训练/学习）	66.6%	T0（dz/log/softmax）	90.7%
T1（矩阵/元素/广播）	5.2%	T1（relu/leaky/convolution）	1.2%
T2（函数/损失/凸函数）	3.8%	T2（batch/miniset/sgd）	1.1%
T3（神经网络/神经元/特征）	3.0%	T3（filter/fl/kernal）	1.0%

续表 1

中文主题&特征词	弹幕占比	英文主题&特征词	弹幕占比
T4（像素/像素点/图像处理）	2.1%	T4（padding/jpg/markdown）	0.9%
T5（方向/计算/阈值）	1.6%	T5（sigmoid/mnist/jio/bv）	0.7%

T6（非线性/函数/线性变换）	1.6%	T6（dropout/value/predict）	0.6%
T7（数据/数据结构/数据库）	1.5%	T7（dont/worry/woory）	0.6%
T8（写错/错误/误导）	1.4%	T8（nlp/idf/tf）	0.4%
T9（过滤器/滤波器/通滤波）	1.3%	T9（bias/high/various）	0.3%
T10（爱猫/分类器/颜色）	1.3%	T10（bug/debug/problem）	0.3%
T11（微积分/高中数学/考研）	1.3%	T11（loss/cost/tradeoff）	0.3%
T12（视频/下个/顺序）	1.2%	T12（cost/function/func）	0.3%
T13（一周/两天/作业）	1.2%	T13（prob/keepprob/best）	0.3%
T14（概率论/条件/概率分布）	1.2%	T14（transformer/vison/transform）	0.2%
T15（四个/三列/三原色）	1.2%	T15（gm/gmdk）	0.2%
T16（代码/编程/编码器）	1.1%	T16（propagation/backgroun）	0.2%
T17（听不懂/不懂装懂/听不出）	1.1%	T17（github/import）	0.2%
T18（算法/优化/贪心）	1.1%	T18（layer/hidden/unit）	0.2%
T19（分子/核是/数量）	1.1%	T19（maxpool/pool/max_pooling）	0.2%

4.1. 主题特征词分布

图 2 为中文弹幕文本主题特征词和权重分布情况，BERTopic 模型输出的高权重特征词能够较好地描述和表达所属主题，例如 Topic1（矩阵计算）、Topic2（损失函数）、Topic3（神经网络）。通过特征词权重也可以帮助明确各主题的核心概念，例如 Topic4（像素）与计算机图像出来有关、Topic7（数据结构）与计算机的专业课《数据结构与算法》有关、Topic9（过滤器）与神经网络设计有关。在线课程为录播课，老师无法和学习者互动，而 Topic0（谢谢老师）主题保留了现实课堂的仪式感，这类弹幕往往重复在视频开头和结尾出现，模拟现实课堂的上下课，让观看者有沉浸式的学习体验。

图 2 *中文主题特征词分布图*

和中文主题不同，英文的主题特征词集中于课程的基础理论和编程实践相关视频，如图 3 中的 Topic1、Topic2 都是和深度学习中的数学理论相关，Topic10 则讨论了编程过程中的 Debug 问题。和中文主题相比，英文各主题的核心概念更加细化，例如 Topic8 在讨论深度学习背景下的自然语言处理方向。英文主题可以帮助学习者获得情感支持，以 Topic7 为例，因为授课教授有一句口头禅“Don’t worry”（别担心），当学习者碰到难点时便会发这句弹幕互相安慰鼓励。这类重复弹幕形成一种“回音室效应”（岳芳等人，2024），仿佛有很多个同伴在陪自己学习，学习者的学习信心受到积极情绪感染开始提升。

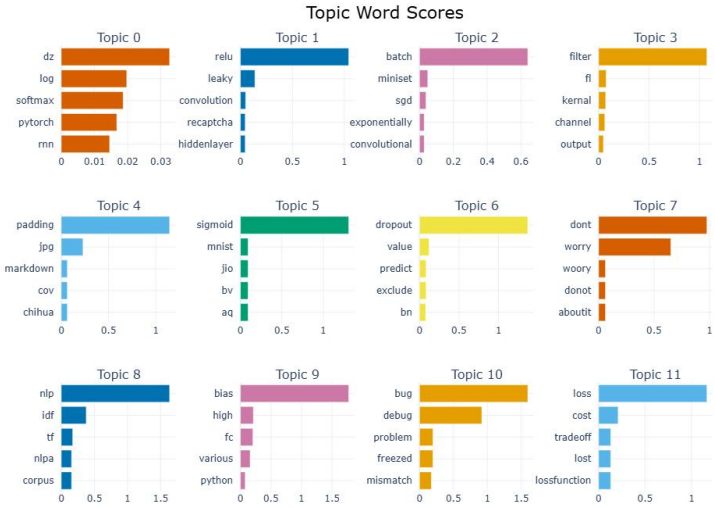

图 3 *英文主题特征词分布图*

4.2. 主题讨论方向识别

根据已识别出主题，基于余弦相似度和层次聚类得到弹幕主要讨论方向。图 4 为中文主题聚类结果。依据聚类结果将 20 个中文主题归纳为计算图像处理、滤波器、编码器、矩阵与神经网络、数据结构与算法、微积分与概率论、损失函数等七个方向。其中矩阵与神经网络、计算图像处理和数据结构与算法等方向的讨论热度较高。在统计学与概率论方向出现了 Topic17（听不懂）主题，由于统计学与概率论属于陈述性知识，内容相对枯燥无聊，因此学习者出现厌学和畏难的负面情绪时会发送消极弹幕，这类弹幕会重复刷屏，导致其他学习者跟风互动，影响学习效果。

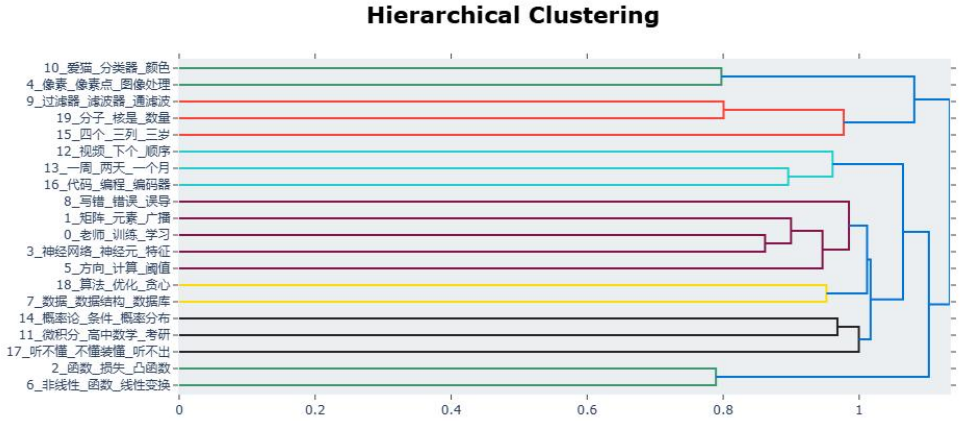

图 4 中文主题层次聚类图

对英文主题的聚类结果如图 5 所示，英文主题可以归纳为损失函数（loss function）、反向传播（Back Propagation）、代码调试（Debug）、深度学习模型框架（TensorFlow）、模型训练（batch）、激活函数（sigmoid）、卷积核心（Kernal）等八个方向。和中文主题相比，英文的主题讨论方向则集中在神经网络的理论学习上，学习者讨论如何基于各类函数的特性，组合设计出最适合目标任务的神经网络框架。代码调试（Debug）方向说明学习者会主动练习课程中的编程任务，并且碰到问题会放到弹幕区讨论，其他人会参与答疑讨论，最后给出解决方案。弹幕交互通过“触发事件-探索-整合-解决”的认知知识构建过程帮助学习者探究知识，理解知识，最终应用知识，增加学习获得感。

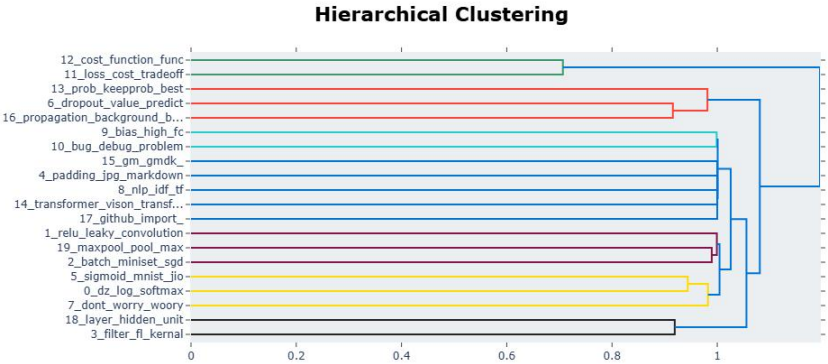

图 5 *英文主题层次聚类图*

4.3. 主题演化趋势分析

BERTopic 的动态主题建模可以直观呈现领域主题的关注热度和学习者讨论度变化情况，便于对深度学习视频的主题演化趋势进行分析。本文以月为单位时间戳切片，将弹幕文本数量视作主题研究热度，分别观察中英文 10 个热门讨论主题的热度变化趋势 。

从图 6 中可以发现所有中文主题的热度在 2021 年 6 月前达到高峰，之后开始滑落。但是在 2022 年 1 月后，热度下降趋势减弱，开始缓慢回升。2023 年和 2024 年的热度变化也基本类似，即 1 月前热度下降，1 月后热度回升。通过对原始弹幕文本进行分析得知，每年 3 月-6 月是高校本科生完成毕业论文的热门时间段，此时会有大量学生来观看学习视频补课（弹幕评论区会有‘毕业论文’相关话题发送），因此弹幕主题热度会出现周期性的波动。2023 年 1 月-6 月，Topic3（神经网络）主题热度快速回升，分析弹幕内容，发现 ChatGPT 相关词频占比超过 0.25%，这说明除了完成毕业论文，仍有学习者是出于对深度学习话题的兴趣来学习课程。

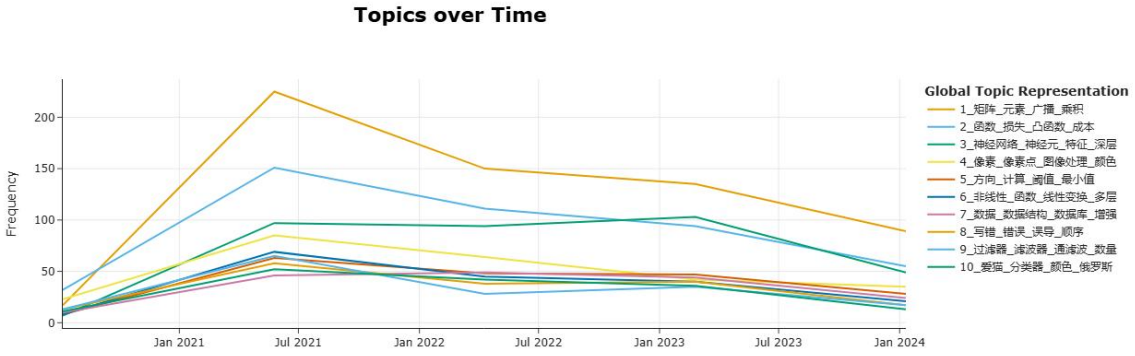

图 6 *中文主题演化趋势图*

英文主题的演化趋势如图 7 所示，与中文一样出现周期性波动。但值得注意的是，2023 开始 Topic3 的热度超过 Topic2，查看 Topic3 相关弹幕文本，发现弹幕从疑惑卷积神经网络中“filter”参数选值开始，不断有学习者通过弹幕对这一概念进行解释，质疑，辩论，并吸引新学习者加入讨论，使得课堂学习不再是老师的单向输出，而是学习者主动参与课堂交互，实现对知识的深度理解。

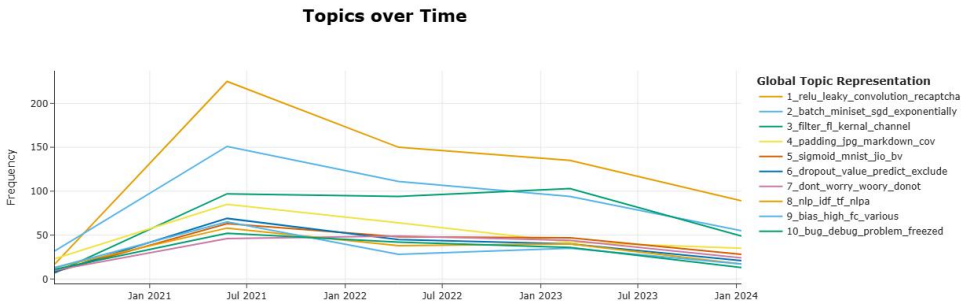

图 7 *英文主题演化趋势图*

5. 总结与讨论

本文使用 Python 程序爬取 B 站平台热门学习类视频弹幕区文本数据，基于 BERTopic 模型挖掘中文和英文弹幕主题的差异性分布、主题讨论方向集合和主题热度演化趋势后，得出如下研究结论：

1.在线课程缺少线下课堂的真实感，学习者容易丢失注意力，而高频的互相问好弹幕维持了在线学习的秩序，让学生意识到在线学习的仪式感，从而加强学习投入。另一方面，本文选择分析的学习视频为英文授课且涉及计算机专业英文单词，有学习者会主动校对字幕错误和解释专业词汇并发送到弹幕区，成为“虚拟”助教，增强在线学习的感染力。

2. 弹幕互动有助于学习者融会贯通所学知识。学习者在完成课堂练习时，会主动发弹幕讨论练习过程中碰到的困难并交流解决方案，最后整合各方发言得出结论，学习者完成了从获取知识到应用知识的构建过程。但是弹幕对于陈述性知识（如数学知识）学习的帮助作用不大，反而部分学习者会因为无法理解知识点频繁发送消极弹幕，削弱其他学习者自信心。

3.弹幕因其短文本和即时性的特征，使得大部分内容属于情感化表达，这恰好满足了学习者的情感支持需求。弹幕交互构建了一个“伪实时”的社交环境，学习者通过弹幕对视频内容展开深入讨论和互动，降低了学习者在线学习的孤独感。弹幕还为学习者提供了跨越时空的生生互动，当碰到学习难点时，不同时间和空间下发送的鼓励加油类弹幕会缓解学习者的紧张情绪。

综上，弹幕可以在一定程度增强在线学习体验，提升学习参与感和获得感，提供情感交互，如果要引入弹幕到教学平台，则需要规范弹幕发送管理机制、调整教学方法。基于本文研究结论，给出以下几点建议：

1.教师应该鼓励学生在遇到知识难点时积极利用弹幕进行讨论和交流，引导学生主动分享疑问并帮助他人解答，增强在线集体学习氛围。

2.平台应优化弹幕功能，提供更灵活的互动形式。例如，设计“即时答疑”弹幕和“知识补充”弹幕，允许学生将自己的解答或建议与他人共享。平台还可以通过数据分析监测弹幕内容，自动过滤无关或负面情绪的弹幕，减少无效或干扰性内容，维护弹幕质量。

3.教师可以定期检查弹幕内容，针对学生的疑问和讨论内容进行总结和补充，进一步促进学生的学习。教师还应注意调整授课节奏，避免因知识内容过于抽象而让学生产生负面情绪。

参考文献

王冲、张雅君和王娟 （2024）。社会大众如何看待生成式人工智能在教育中的应用？——对 B 站 ChatGPT 话题弹幕文本的舆情主题与情感分析。图书馆论坛（10）, 61-71。

王歌 （2024）。基于视频弹幕的情感分析与动态主题挖掘研究。硕士学位论文。西南财经大学。

赖显静（2023）。基于 LDA 主题模型的 MOOC 课程评论文本分析。现代信息科技(04), 43-46。doi:10.19850/j.cnki.2096-4706.2023.04.011。
刘三女牙、彭晛、刘智、孙建文和刘海 （2017）。面向 MOOC 课程评论的学习者话题挖掘研究。电化教育研究（10）, 30-36。doi:10.13811/j.cnki.eer.2017.10.005。
易田 （2018）。在线教学视频中弹幕对不同类型知识学习的影响（硕士学位论文，华中师范大学）。
陈忆金、卓林锴和赵一鸣 （2021）。学习类视频弹幕用户的交互行为研究。图书馆论坛（09), 95-101+124。
杨思洛和于永浩 （2024）。基于 BERTopic 模型的国内信息资源管理研究主题挖掘与演化分析。情报科学。
岳芳、樊茂瑞、高子雅、郭剑锋和肖吉军 （2024）。开放式知识协同平台中的“回音室效应”研究——以 Bilibili“新能源汽车”视频评论为例。情报理论与实践（02）, 124-131。doi:10.16353/j.cnki.1000-7490.2024.02.017。
张婧婧、杨业宏和安欣（2017）。弹幕视频中的学习交互分析。中国远程教育(11), 22-30+79-80。doi:10.13541/j.cnki.chinade.20171120.008。
张文兰、陈力行和孙梦洋 （2022）。弹幕交互为大学生在线学习带来了什么？——基于扎根理论的质性分析。现代远距离教育(05), 12-19。doi:10.13927/j.cnki.yuan.20220706.007。
周璐（2024）。弹幕背后的探究社区：初中生在线学习互动行为分析。少年儿童研究（05）, 11-19。doi:CNKI:SUN:SNEY.0.2024-05-002。
穆肃、王孝金、冯冠朝 & 张晗 （2020）。在线同步教学中交互的设计与实施。中国电化教育（11）, 52-59+66。doi:CNKI:SUN:ZDJY.0.2020-11-007。
田园和宫婷婷 （2021）。基于 LDA 模型的在线教学需求数据主题挖掘研究。情报科学（09), 110-116。doi:10.13833/j.issn.1007-7634.2021.09.015。
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in sociology, 7, 886498.
Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
Lin, X., Huang, M., & Cordie, L. (2018). An exploratory study: using Danmaku in online video-based lectures. Educational Media International, 55(3), 273-286.