基于多模态的涉藏媒体短视频舆情感知

谢森元　郑欣宇　胡子奇　满伟栋　冯怡宁
（西藏大学 信息科学技术学院，西藏　拉萨　850000）

【摘　要】与以往单模态相比，多模态借助对混合数据各个模态的信息予以处理，使得具备相似语义的模态具有相似的结果，通过映射和处理单模态的基础上实现语义一致。文中从视觉、听觉、文本处理涉藏媒体短视频的舆情感知，具体而言使用Vision Transformer（VIT）、Transformer端对端方式进行多模态融合，提出模型构建方法和工作方法。

【关键词】多模态；舆情感知；Transformer；数据集搭建；深度学习；数据集搭建

一、涉藏社会媒体短视频舆情感知的背景和意义

（一）研究背景

舆情是在一定时期，一定社会空间内民意的一种综合反映[1]，是群体性的思想、情绪、意见和态度等的总和。舆情的发展态势与特征使得社会取得一定服务。很多学者致力于研究如何构建良好的舆论环境保障民众的信息安全与健康。

西藏是重要的国家安全屏障和生态安全屏障，是维护祖国统一、贯彻新时代社会主义的重点地区。西藏现取得“十三五”时期经济社会发展取得的全方位进步、历史性成就，准确把握“十四五”时期经济社会发展新形势的重要时期。当前，面对世界百年未有之大变局、中华民族伟大复兴战略全局。现在网络社交媒体发达，各种舆论涌现，网络上出现不当言论产生舆情，影响社会稳定发展。

（二）研究意义

近年，随着移动互联网的快速发展，上网人数呈现巨额增长。藏族地区的信息化、数字化程度也随着移动互联网的快速发展日趋成熟。在广泛传播的大数据下，海量的信息伴随而来。互联网科技的迅猛发展，使得社交媒体成为人们获取文本、音频、视频等信息、表达观点和情感交流的重要平台，其中包括使用藏文的社交媒体。随着社交媒体的普及和使用用户数量的增加，社交媒体上的信息情感的感知和情感识别的研究变得越来越重要。互联网在网络环境安全方面起到大额占比。社交媒体上的不良信息在信息化时代产生的危害将会成为藏区网络的风险。因此增强社交媒体上藏文信息的治理手段刻不容缓，这对维护社会和谐，以智能化引领藏区发展，帮助社会维持长足经济发展有着重要意义。

由于互联网的虚拟性和部分网民的非理性因素，部分信息被歪曲或编造，并且迅速传播。 藏区网络舆情主要分为形成阶段和发酵阶段，并与不稳定事件产生相互影响、相互推动的效果。2023年9月，西藏自治区互联网违法和不良信息举报中心通过网站、微信、微博、电话和APP等渠道接收举报信息10259条，受理有效举报信息3855条。接收举报信息与2022年同期相比，有效信息增长159.4%。受理的举报信息均已依法依规研判，分类协调处置。因此参与网络综合治理的积极性、主动性日益提高。本月受理的各类有效举报中，涉政治类有害信息1253条，占32.5%；涉民生诉求类信息973条，占25.2%；涉淫秽色情类有害信息594条，占15.4%；其他类有害信息1035条，占26.9%。

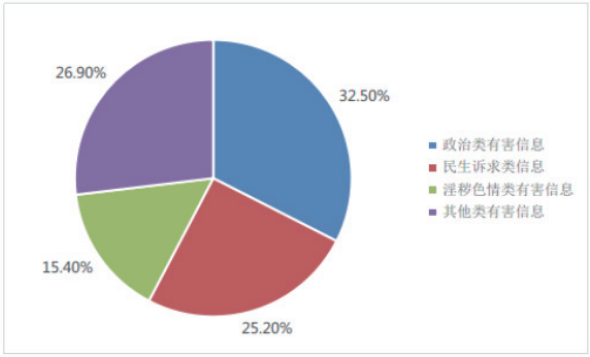

图1 2023年9月有害信息占比图

同时，目前对于藏区社交媒体的信息感知和舆情智能分析的研究也相对较少。主要因为藏文是一种特殊的语言，其文本、音频和视频的特点以及情感表达方式与其他语言存在明显的差异。因此，针对社交媒体的信息感知和涉藏舆情智能分析技术的提出不仅有需要，而且可以有效降低不良信息传播风险，有助于保护公众免受虚假信息

和其他社会负面信息的影响，营造健康网络生态和清朗网络空间。

二、研究内容概述

文章的研究目标是基于多模态信息整合，运用深度学习的方法，对社交媒体中的短视频进行分类及对涉藏舆情内容进行智能分析研究。

文章是基于西藏地区的研究，因为语言环境的不同，内地和西藏地区的短视频舆情感知最大差异是在于音频部分，因此需要单独训练藏语翻译的音频模型数据集，构建数据集的主要点是在音频数据集标注，音频数据集标注完成后使用精确对应音频内容的文本数据库记录音频内容，将文本数据库以label标签标注0或1用于区分舆情状态，使模型通过识别文本进行舆情感知。视频数据集通过标注label后进行视频patch识别进行舆情感知。

（一）藏文音频数据集和视频数据集的构建

针对涉藏舆情多模态感知问题的不断深入，目前许多数据集大部分是单模态的，单一的从文本，语音，视频收集数据，而多模态的数据集相对较少。在生成对抗网络技术的不断发展和完善下，Deep Fake等深度伪造技术也在不断进步。好的数据集是研究的关键，因此数据集的质量一定程度上影响着整个模型的效果。Heo 等人提出了一种用于深度伪造检测的高效视觉 transformer 模型来同时提取局部和全局特征，其结合CNN特征和基于块的定位与所有位置进行交互来指定伪影区域并添 加了蒸馏token，通过sigmoid函数使用二进制交叉熵训练logit。

音频数据集的搭建 需要将非藏语的音频文件使用人工翻译的方式翻译为以藏文为基准的音频数据，之后使用音频标注软件将藏文发音细致标注。以此完成藏文音频数据集的构成和搭建。

视频数据集的搭建需先以翻译好音频数据调整视频长度，其次使用视频标注软件对视频舆情内容进行label标签，之后经过稀疏时间算法减小视频计算量。以此完成视频数据集的收集。

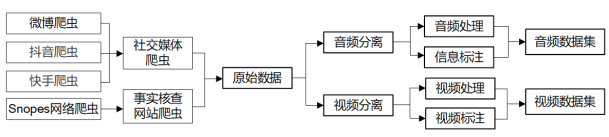

图2 数据集搭建过程

（二）视频模态的特征处理与提取

使用密集采样的方式会导致帧之间存在冗余信息，因此现如今的特征提取方法主要依赖于光流图像的提取，采用稀疏时间采样策略一定程度上能消除冗余信息，减少视频图片的计算量，同时能简化算法复杂度。研究比较基于卷积网络的特征提取和基于transformer的特征提取两者的共同点与差异点，综合考虑全局信息和局部信息，设计选取最佳的提取特征的网络或者方法。Vision Transformer[1]为视觉领域提供了强大的端到端解决方案，这是一个基于标准Transformer 编码器的面向图像的网络。它有一个特定于图像的输入pipeline，其中输入图像必须分成固定大小的patch。将标注好的视频稀疏时间采样后逐帧转化为图片存放在文件夹内，保证视频文件夹与音频文件夹中的标注数据顺序一致，模型读取已完成标注的视频后提取视频特征。

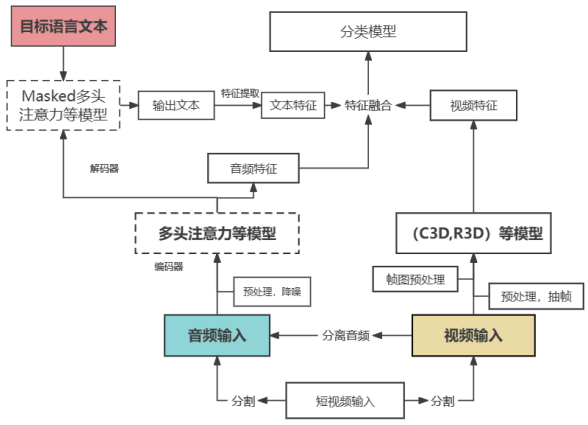

图3 对音频加入端对端模型的多流架构图

（三）多模态的特征提取与特征融合

多模态信息的特征使用简单的融合或者拼接会丢失各个模态之的隐秘联系，采用基于低秩多模态融合的网络Low-rank Multimodal Fusion（LMF）进行改进，同时引入多头注意力机制来训练得到不同模态在融合特征中的注意力权重，结合特征聚类算法对帧级特征聚合成视频级特征，并采用压缩激励全局平均池化方法来激励由有用的特征，抑制无关的信息。将权重分解为低阶因子，这样可以减少模型中参数的数量，通过利用低阶权重张量和输入张量的并行分解来有效地进行基于张量的融合Attention 机制，实质上是一个携带权值的寻址过程。模型对帧级特征进行主成分分析降维后应用，利于下一层模型的泛化性能。

（四）藏文媒体、舆情倾向性分析研究

将上述各个单模态得到的特征输入多模态的全连接层中，得到最后的分类结果。对藏文媒体短视频、舆情进行倾向性分析，以正面、负面及中性对

藏文新闻的情感性进行多种划分。

三、研究思路和方法

（一）研究思路

1.通过Python来爬取目标新闻媒体网站（如微博，抖音，快手等）的藏语短视频。2.对爬取的音视频材料进行预处理，将音视频分割，分别构建藏文媒体音频、视频数据集，并对每项数据进行分类，标注发布者及真实性，此数据集将用来训练网络模型。3.根据单帧图像和光流信息作为提取视频特征的基本信息，在此基础之上，运用空间注意力机制的残差网络提取视频在的动作序列空间特征，基于帧级处理融合框架来探寻各帧之间的相关联系，使用基于Transformer的分类模型Timesformer，将自注意力扩展到时空的3D空间。4.首先将输入的音频数据集进行预处理，提高特征提取的效果，然后进入基于Transformer的端对端模型的编码器，经过多头注意力，前馈网络，残差连接，层归一化等后，进入编码器，最后生成目标文本。5.提出基于强化学习的搜索空间来寻找最优的融合网络架构，基于线性池融合的方法和基于自注意力融合的方法来找到多模态特征之间的联系，然后通过改进Bert模型来训练参数，最后得出分类结果。

（二）工作方法

1.音频数据在模型中的转换

端对端简易而言是使用模型完成语音识别的整个过程，直接以语音为输入、文本为输出的整体过程。模型将标注完成的语音首先转化成对应的文字输出，其次将识别转化的文字做概率预测分析文本情感。

2.视频数据在模型中的转换

标注完成的视频数据经过patch分割识别图片特征后，输出概率预测参与多模态融合的步骤。

3.多模态融合中数据的转换

使用正则化术语实现多模态共享语义捕获，文献[2]将多模态融合方法分为模型无关方法和基于模型方法，使用较复杂的混合融合方法解决涉藏社会多媒体舆情感知。文献[3]使用MKL进行多模态融合，使用傅里叶变换计算映射函数，取得较好的融合效果。

文本使用Low-rank Multimodal Fusion（LMF）[4]进行多模态融合，不仅降低参数量，同时提升了计算速度，将视频模态和音频文本模态并行提取各模态特征，使用linear后融合成新的特征，最后输出多模态的预测结果。

四、方法总结

单模态模型主要通过单一线性或非线性进行映射后，得到模态的语义特征。相比之下，多模态模型通过模态之间的相关性，不仅保留原有单模态特征的基础上，增加了特征维度，避免了单一维度产生的过拟合性，而且使用Low-rank Multimodal Fusion（LMF）减低了多模态的计算量，从视频、音频、文本方面进行数据收集，数据归类，数据分析。

五、结束语

多模态涉藏多媒体舆情感知一定程度上加强了网络安全，使用深度学习多模态技术缓解了部分网络安全中心的压力，可以使机器自主监督藏语语音的多媒体短视频舆情，净化网络方面做出一定成果与贡献。

在如今已有的领域内综合使用各模态之间的数据收集、搭建以及模态融合的方式进行了舆情管理。使用低秩方式验证了低秩的小运算高精度特点。但如何满足多模态学习需要的庞大数据集依旧是一个难题，深度学习模型对数据需求量非常之高。研究中依旧是数据集遇到的难题，藏语的口音问题，视频语音的翻译问题都会使得模型在实际应用中产生一定的偏差。因此，自建高质量数据集是模型重要的一个环节。

参考文献：

[1]余才忠，熊峰，陈慧芳.舆情民意与司法公正——网络环境下司法舆情的特点及应对[J].法制与社会：旬刊，2011（12）：2.

[2]Dosovitskiy A，Beyer L，Kolesnikov A，et al.An Image is Worth 16x16 Words：Transformers for Image Recognition at Scale[C]//International Conference on Learning Representations.2021.

[3]何俊，张彩庆，李小珍，等.面向深度学习的多模态融合技术研究综述[J].计算机工程，2020，46（05）：1-11.

[4]Natasha Jaques，Taylor S，Akane Sano，et al.Multi-task，Multi-Kernel Learning for Estimating Individual Wellbeing[J].[2024-06-27].

[5]Liu Z，Shen Y，Lakshminarasimhan V B，et al.Efficient Low-rank Multimodal Fusion with Modality-Specific Factors[J]. 2018.

作者简介：

谢森元（2002—），男，汉族，甘肃兰州人，国家级实验室管理员，本科，研究方向：深度学习。