# 基于多模态数据的社交平台谣言检测

**教学适用的课程：**《大数据架构与技术》

**教学适用的专业：**电子信息

**教学目的与用途：**学习实际社交平台谣言检测中的大数据应用；掌握大数据处理与分析的基本流程；掌握机器学习、深度学习算法在大数据系统中的应用。

**教学内容：**

（1）理论依据：本案例基于大数据技术对社交媒体平台中的多模态推文进行谣言监测。案例整合了多种的数据挖掘与深度学习方法，进行数据挖掘与基于多种机器学习、深度学习算法的社交平台谣言监测。

（2）涉及知识点：社交媒体大数据、多模态学习、自然语言处理、分类算法。

（3）分析路径：首先介绍社交平台谣言检测的背景，并进行数据集展示，介绍相关数据形式。然后学习本案例中给出的多种机器学习算法与深度学习算法，理解不同文本嵌入方式以及多模态数据融合的方式。最后，使用二分类问题常用的评价指标，对各种算法进行对比与分析。

**启发思考题：**

（1）本文使用了预训练的Bert与VGG模型提取特征，且没有对特征提取网络进行微调，但是深度学习算法的源域与目标域必然存在不同程度的差别，对预训练网络进行微调可以有效提升特征的有效性。但是大型网络在小规模数据集上训练易发生梯度爆炸、梯度消失、过拟合等问题，能否克服这些问题提升识别效果。

（2）本文将文本特征与图像特征拼接直接后送至全连接网络进行分类，能否寻找其它方法进一步考虑两种特征之间的交互，以及动态决定单模态特征对整体分类的作用等。

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常见的自然语言处理、图像特征提取、机器学习与深度学习算法的理论知识。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：在当今的移动互联网时代，微博、今日头条及各大新闻类APP成为了信息传播的重要途径，为人们的生活带来了很大的便利。与此同时，虚假新闻与网络谣言也大量地充斥于互联网上，其无休止地传播会为人们的认知甚至社会的稳定都带来很大危害。谣言或虚假新闻检测的目标为识别出捏造的、可以被证实为虚假的新闻内容。这节课我们将接触一系列基于大数据与学习算法的社交平台中谣言检测实践。
5. 结束总结：本科陈跟我们选取基于多模态数据的社交平台谣言检测为主要问题，使用来源于微博平台的大量谣言与非谣言数据，结合数据挖掘、机器学习、深度学习的相关方法，进行谣言检测。需要注意的是，课程所学内容是有限的，希望对该领域感兴趣的同学可以自行搜集相关资料，进一步了解大数据在社交平台谣言检测的其它方法及相关技术。
6. 案例引导建议：在教师简要分析应用场景后，鼓励学生自主考虑如何使用所学知识解决该场景问题。

**参考文献：**

[1] Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM international conference on Multimedia. 2017: 795-816.

[2] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[3] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[4] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.

**其他教学支持材料：**

1. 本案例提供配套的PPT、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），Pytorch（深度学习算法库），Gensim（自然语言处理）。