探索预训练模型
在新闻标题聚类中的应用

1024040824 孙士宇

2025.6.14

content

目录

引言与背景

海量信息涌现

互联网与社交媒体的兴起导致新闻数据呈爆炸式增长，每秒产生前所未有的信息量，传统人工分类难以应对。

效率与成本问题

依赖人力进行新闻文章分类不仅效率低下、成本高昂，还可能因主观因素引入偏见，影响分类准确性。

实时性需求

实时生成的大量新闻数据要求快速响应与处理机制，传统方法难以满足现代信息管理的时效性要求。

引言与背景

BERT的革新

BERT通过双向Transformer架构，实现了对输入序列的深度理解，显著提升了NLP任务的处理能力。

自注意力机制

Transformer引入的自注意力机制，使模型能同时考虑输入序列的所有位置，捕捉更复杂的语义关系。

预训练与微调

BERT采用两阶段策略：先在大量无标注文本上预训练，再针对具体任务进行微调，有效利用语言的通用特征。

多头注意力

通过多头注意力，BERT能在并行处理不同信息子空间的同时，增强模型的表达能力和计算效率。

方法论与实验设计

BERT嵌入转换

利用BERT将新闻标题转换为语义丰富的文档嵌入，捕捉文本的上下文意义，为下游聚类算法提供高效处理的数据形式。

K-Means聚类执行

基于BERT生成的嵌入向量，运用K-Means算法进行聚类，通过迭代优化过程，实现对新闻文本的自动分类。

评价指标应用

采用Silhouette系数、Calinski-Harabasz指数和Davies-Bouldin指数等指标，评估聚类结果的质量和有效性。

框架整合流程

整合BERT的文本理解能力和K-Means的聚类效率，构建一个高效、可扩展的新闻文本聚类框架，应对高维文本数据的挑战。

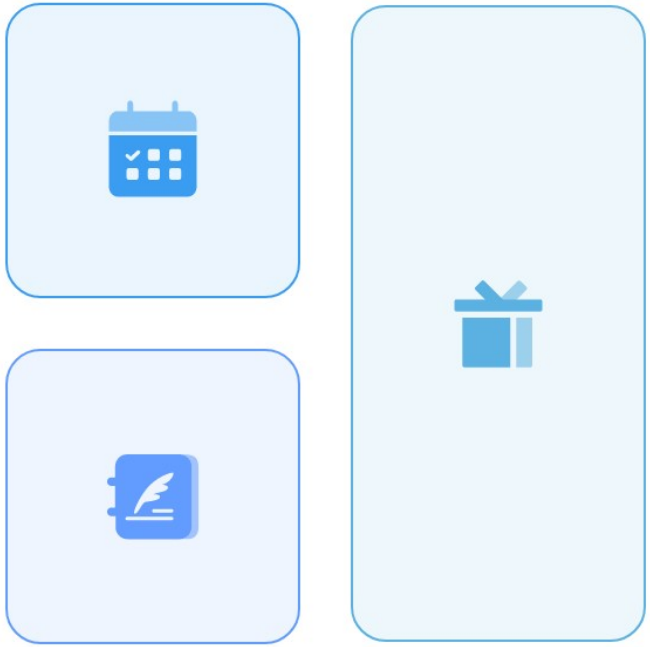
方法论与实验设计

PCA线性降维

通过计算协方差矩阵和特征向量，PCA将高维数据投影到低维空间，保留最大方差方向。适用于去除冗余信息，提高计算效率。

UMAP非线性映射

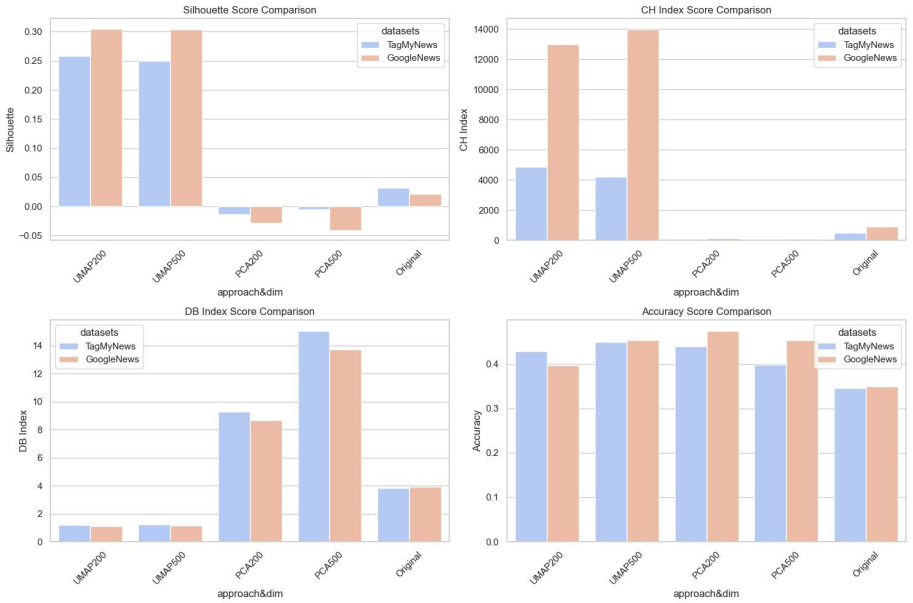
基于流形学习理论，UMAP优化局部与全局数据结构，有效揭示复杂关系和潜在聚类结构，尤其适合高维数据可视化。



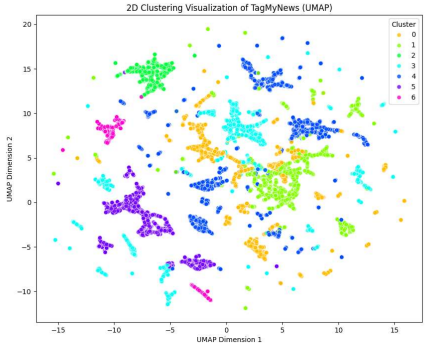
技术对比与选择

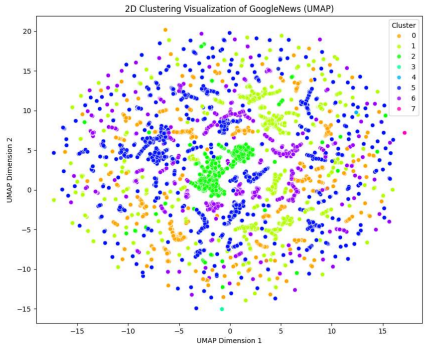
PCA擅长线性结构，而UMAP能捕捉非线性关系，两者可优化聚类效果，提升模型性能。

实验结果与分析



指标概览：评估采用Silhouette Score、Calinski-Harabasz Index、Davies-Bouldin Index及Accuracy四大指标





UMAP vs PCA：UMAP在多数指标上超越PCA，尤其在Silhouette Score与Calinski-Harabasz Index表现突出，揭示复杂非线性关系

局限性探讨：尽管UMAP增强聚类效果，但可视化结果不够清晰，提示高维数据降维可能破坏原有结构，需谨慎选择降维策略。

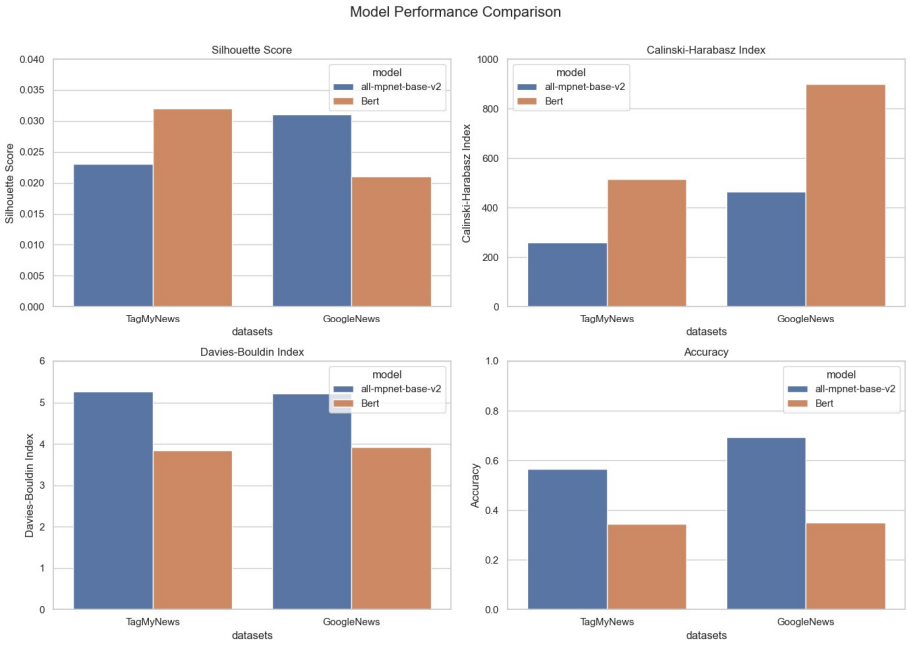
模型微调与性能比较

微调模型优势

微调模型如all-mpnet-base-v2在分类准确性上表现更佳，尤其在GoogleNews数据集上，展现出对特定任务的适应性。

BERT性能稳定

BERT在Calinski-Harabasz Index和Silhouette系数指标上保持稳定表现，证明其在聚类有效性上的可靠性。



任务导向选择：基于具体任务需求选择模型至关重要，all-mpnet-base-v2在分类精度上领先，而BERT则在聚类质量上更具优势。

结论与未来方向

预训练模型潜力

预训练模型如BERT结合K-Means展现了处理大规模新闻数据的潜力，有效捕捉新闻标题的语义特征。

维度缩减提升

UMAP在维度缩减方面优于PCA，显著提升了聚类性能，尤其是在Silhouette Score和Calinski-Harabasz Index指标上。

微调模型优势

微调模型如all-mpnet-base-v2在分类准确性上表现突出，特别是在GoogleNews数据集上，表明其对特定任务的适应性。

结论与未来方向

模型微调策略

深入探索模型微调策略，针对特定领域数据进行优化，以提升模型对特定语境的理解能力，进一步提高聚类精度。

算法创新

研究更先进的聚类算法，如层次聚类或DBSCAN，结合深度学习技术，以增强模型处理复杂数据结构的能力。

多模态信息融合

整合图像、音频等多模态信息，与文本数据相结合，构建更全面的新闻事件表示，提升聚类的多样性和准确性。

谢 谢 观 看