ClusTop：一种无监督的一体化文本聚类与主题提取框架

陈中涛[a,1]，米成虎[a,1]，SiweiDuo[b,2]，何景飞[a]，以及YatongZhou[aa]河北工业大学电子与信息工程学院[b]天津汇智星源信息技术有限公司

arXiv:2301.00818v1 [cs.CL] 3 Jan 2023

摘要

文本聚类和主题提取是文本挖掘中的两项重要任务。通常，这两项任务是分开执行的。为了通过主题提取促进聚类，我们可以先将文本投影到主题空间中，再使用聚类算法获得簇。为了通过聚类促进主题提取，我们可以先用聚类算法获得簇，再提取簇特定主题。然而，这种简单策略忽略了一个事实：文本聚类和主题提取密切相关，二者存在先有鸡还是先有蛋的关系。将它们分开执行，无法使二者相互促进，从而达到最佳的整体性能。本文提出了一种无监督文本聚类与主题提取框架（ClusTop），将文本聚类和主题提取整合到统一框架中，能够同时实现高质量聚类并从每个簇中提取主题。我们的框架包含四个组成部分：增强语言模型训练、降维、聚类和主题提取。其中，增强语言模型可视为连接聚类与主题提取的桥梁。一方面，它生成具有强簇结构的文本嵌入，有助于有效开展文本聚类；另一方面，由于采用了自注意力架构，它会高度关注与主题提取相关的词语。此外，增强语言模型的训练采用无监督方式。在两个数据集上进行的实验验证了我们框架的有效性，并为该框架中的不同模型组合提供了基准。

关键词： 文本聚类、主题提取、自注意力、预训练语言模型。

[1]前两位作者贡献相同（按字母顺序排列）。[2]通讯作者：
siweiduo@cispirit.com

1 引言

随着信息技术的快速发展，每天都会产生海量数据。自动文本聚类和主题提取具有直观的重要性，因为它们能够促进许多网络应用的发展，例如问答系统 [1], 舆情监测 [2],和个性化推荐 [3]。文本聚类主要包含两个组成部分：文本表示和聚类。现有的文本聚类方法主要专注于优化文本表示模型和聚类算法，以获得更好的聚类结果。

文本表示模型通过捕捉语义信息，将目标文本转换为向量表示，以便聚类算法进行处理。近年来，预训练语言模型（如 BERT [4]）在生成上下文词嵌入和句子嵌入方面表现出色。然而，基于 BERT 的模型会使句子形成各向异性语义空间 [5], ，这会损害其在语义相似度相关任务（如文本聚类）中的表现。为缓解这一问题，BERT-flow [5] 和 BERT-whitening[6] 通过将获得的嵌入映射到各向同性分布来进行后处理。SimCSE [7], 、ConSERT [8], 、ESimCSE [9]和 DiffCSE [10] 提出了对比学习框架，以消除句子嵌入的各向异性。尽管上述方法已被证明在句子语义相似度任务中有效，但它们并不适合直接用于文本聚类，因为同一簇中含义不同的句子会被过度区分。同一簇中句子之间的距离甚至可能大于不同簇中句子之间的距离，而且句子的整体分布无法呈现清晰的簇结构，这不利于后续聚类算法的应用。为此，我们提出一种无监督方法，用于获得增强语言模型，使其能够生成簇结构分离良好且均衡的句子嵌入，从而更适用于聚类任务。具体而言，我们使用伪标注数据，在分类任务上对预训练语言模型进行微调，以获得增强语言模型。伪标注数据通过一种基于聚类的伪标签方法获得：首先使用预训练语言模型生成句子嵌入，然后应用 UMAP（统一流形逼近与投影） [11]进行降维，并采用基于密度的聚类算法（如 HDBSCAN（具有噪声的分层密度聚类） [12], ）筛选高密度区域中的非噪声数据。

由于增强语言模型生成的句子嵌入具有高维特性，基于欧几里得距离的聚类算法会受到维数灾难的影响而失效。一种直接的方法是降低嵌入的维度。

现有的降维技术可分为两类：线性[13, 14, 15]和非线性 [16, 17, 18, 19], ， 其中 PCA [13]和 t-SNE [19]是两种广为人知的方法。UMAP [11] 已被证明比 t-SNE 能捕捉更多的全局结构，并且具有更优的运行时性能。此外，作为一种流形近似与投影方法，UMAP 具有坚实的理论基础。由于簇的数量未知，且待聚类的文本中可能包含噪声点，为避免将这些点分配到不相关的簇中，最好选择 DBSCAN [20]和 HDBSCAN [12]等聚类算法。在我们的实验中，我们还通过为 K-Means 指定一个合理的簇数量，将这一广为人知的基于划分的聚类算法纳入比较。

对于聚类文本，可以使用传统的关键词提取方法词频-逆文档频率 [21, 22]和 TextRank [23] 来提取簇特定主题。词频-逆文档频率易于实现，但无法捕捉语义，且严重依赖语料库和逆文档频率的取值。TextRank 的计算复杂度较高，并且会忽略出现概率较低、但在上下文中具有重要意义的关键词。无论聚类结果如何，都可以独立使用传统主题模型 LDA [24] 提取主题，但它需要预先确定主题数量，并受到其对数据所作假设的限制。利用预训练语言模型学习到的语义信息，keyBERT [25] 借助来自 BERT 的嵌入，找出与文档最相似的关键词。然而，由于 BERT 接受的最大长度有限（不超过 512 个词元），只能从给定簇中选取部分句子进行拼接，直至达到总长度限制，这无疑会导致准确率下降。

为了获取能够描述每个簇主要主题的关键词，我们提出了一种基于注意力的主题提取方法。我们方法的核心是充分发挥增强语言模型的潜力，该模型经过训练，具备强大的分类能力。一方面，它可以提供包含显著类别特征的句子嵌入，使聚类更加容易；另一方面，语言模型最重要的组成部分——自注意力机制——能够告诉我们句子中的哪些词元发挥着重要作用，从而有助于主题提取任务。基于一种直观的认识，即一个词元在句子中越重要，其他词元对它的注意力就越高，我们的方法通过计算句子中所有词元对给定词元的注意力权重之和来衡量该词元在句子中的重要性，进而获得每个簇的主题词。值得注意的是，增强语言模型能够自然地将聚类和主题提取过程整合为一体。

我们的主要贡献可概括如下：

1）我们提出了一个无监督的集成式文本聚类与主题提取框架 ClusTop，该框架由增强语言模型训练、降维、聚类和基于注意力的主题提取组成。在两个数据集上的实验表明，我们的框架具有良好的有效性。此外，我们还为不同的模型组合提供了基准。

2）我们提出了一种无监督方法，用于获得一种增强语言模型，该模型能够生成具有鲜明且均衡簇形状的句子表示，使其更适用于后续聚类。

3）我们提出了一种基于注意力的主题提取方法，用于获取每个簇的关键词。由于挖掘出了自注意力机制的潜力，所获得的关键词与簇语义更加接近。

本文组织结构如下：第二节讨论相关工作。第三节详细介绍所提出的文本聚类与主题提取框架。第四节展示实验结果，结论将在下一节给出。

2 相关工作

文本聚类和主题提取是文本挖掘领域中两个相互关联的任务。先对文本进行聚类，再为每个类别提取关键词，这已成为最具代表性的主题提取框架之一。据我们所知，基于这一框架取得的最先进成果是 Top2Vec [26] 以及近期提出的 BERTopic [27]。这两项工作最初都是为主题提取而开发的，但都以聚类为基础。具体而言，它们都先通过嵌入、降维和聚类流程对文本进行聚类，然后提取每个簇的关键词，作为其主题的描述。在文本嵌入步骤中，Top2Vec 基于 doc2vec [28]，而 BERTopic 基于预训练语言模型；二者分别使用 UMAP 和 HDBSCAN 进行降维和聚类。这两种方法的主要区别在于如何从每个簇中提取关键词。为获取每个簇的主题词，Top2Vec 将簇质心作为主题向量，并认为词嵌入越接近主题向量，就越能代表该主题。然而，聚类与主题词选择之间存在明显的不兼容：聚类依据的是数据分布的密度，而词语选择依据的是与簇质心之间的距离。对于条带状或形状不规则的簇，这种方法容易误选其他簇中的词语。缩小词语选择半径是解决上述问题最直观的方法，但这样会漏掉许多候选词。

BERTopic 提出了一种基于类别的 TF-IDF 变体（c-TF-IDF），用于从每个簇中提取主题词，从而解决了质心相关问题。然而，由于其具有 TF-IDF 的性质，仍会受到词频和簇数量的显著影响。值得注意的是，这两种方法都没有利用预训练语言模型的分布式表示；与基于频率的统计方法相比，这类表示能够更好地捕捉语义。

我们的框架在文本嵌入和主题提取这两个方面与上述两种方法有着本质区别。在文本嵌入方面，我们提出的方法并未直接使用预训练语言模型，而是以簇增强语言模型（即类别表征能力强的语言模型）取而代之，并通过无监督方式训练该增强语言模型。具体而言，原始预训练语言模型生成的嵌入在向量空间中的分布更加均匀，簇结构不明显。直接使用 HDBSCAN 进行聚类会产生大量噪声点，导致结果的召回率较低。相比之下，增强语言模型生成的文本嵌入具有鲜明的簇结构，簇内点分布更加紧密，不同簇之间的距离更大，边界也更加清晰。这无疑有利于后续使用 HDBSCAN 进行聚类，从而减少噪声点数量并提高召回率。在主题提取方面，我们的方法通过将增强语言模型的自注意力机制直接用于主题提取任务，最大限度地发挥其优势。鉴于增强语言模型能够取得良好的文本聚类结果，可以看出它已经具备较强的类别区分能力。从句子级视角来看，增强语言模型提供的嵌入具有更鲜明的簇形状，有利于聚类；从词元级视角来看，它会将注意力转向与簇主题最相关的词元，从而自然地为每个簇提供最贴切的主题词。与上述方法相比，我们的方法利用增强语言模型将聚类和主题提取衔接起来，最大限度地发挥了语言模型的语义表示能力。

3 框架

我们的文本聚类和主题提取框架主要分为两部分：基于增强语言模型的文本聚类，以及基于注意力的主题提取，如图 1 所示。首先，需要获得一个增强语言模型，用于文本表示。增强语言模型可以视为对原始预训练语言模型的改进，它能够扩大簇之间的距离，并使簇内的点更加紧凑，从而使生成的文本嵌入更适合聚类任务。

由于得到的文本嵌入维度较高，因此在聚类前需要进行降维。最后，应用基于注意力的主题提取方法，从每个簇中提取主题。接下来，我们将详细介绍文本聚类和主题提取框架，重点说明如何获得增强语言模型，以及如何基于注意力机制提取主题。

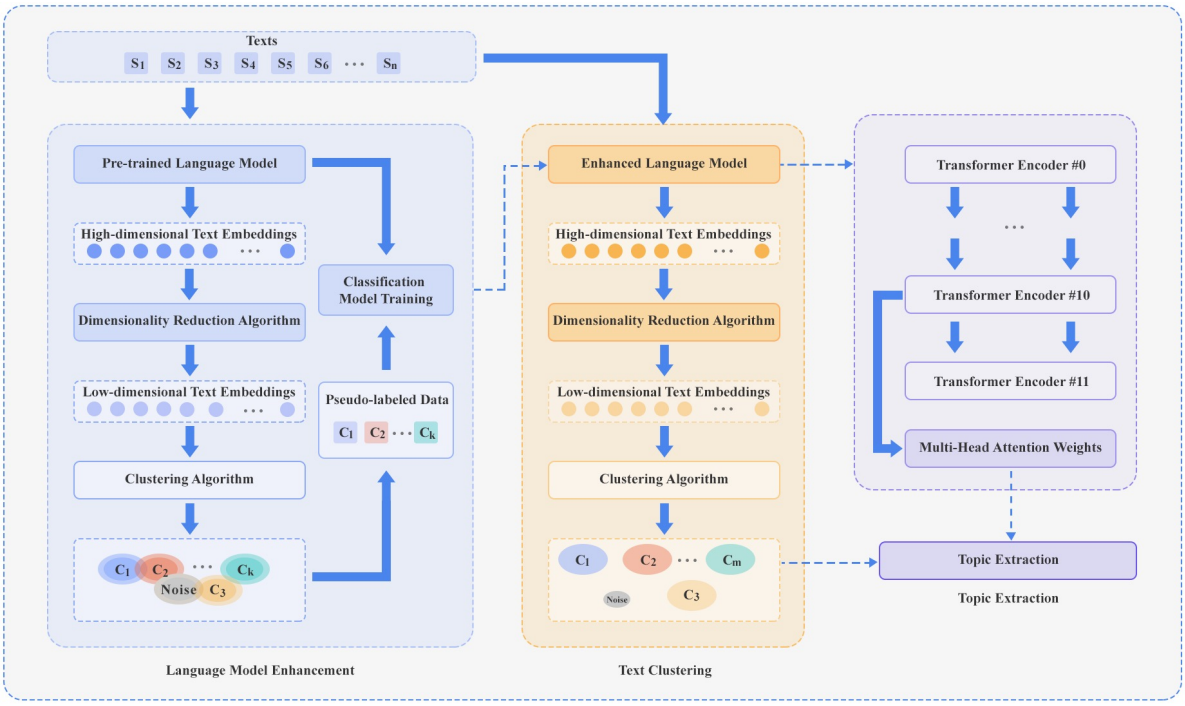


图 1：框架的第一部分是获取具有强簇特征的增强语言模型；第二部分包括基于增强语言模型的文本表示、降维和聚类；第三部分是基于注意力的主题提取。

3.1 基于增强语言模型的文本聚类

基于Transformer的预训练语言模型（未经微调）具备一定的语义表示能力，但其性能尚不足以直接用于文本聚类，因为文本嵌入的分布没有明显的簇结构，因而导致聚类结果的准确率较低。具体而言，由于缺乏清晰的簇结构，基于划分的聚类方法会因不同簇之间边界不规则而错误分类数据点；基于密度的聚类方法则会产生大量本应属于某些簇的噪声点。受微调范式的启发，为解决上述问题，我们在文本分类任务上训练语言模型，使其学习簇级语义，从而让增强语言模型对句子的簇特征更加敏感。

将聚类特征增强语言模型应用于文本嵌入，可得到具有清晰簇结构的文本表示空间，从而更有利于聚类。本节包含三个步骤：训练聚类特征增强语言模型、降维和聚类。我们将详细介绍增强语言模型的训练过程，这也是本研究的主要贡献之一。

增强语言模型的训练

受先预训练再微调范式的启发，我们通过训练分类任务，使语言模型学习聚类层级的语义信息，从而调整原始文本表示的分布，使其具有更强的聚类特征，同一簇内的点更加紧密，不同簇之间的边界更加清晰。为了获得用于训练分类模型的标注数据，我们提出了一个无监督框架，可分为四个步骤：文本表示、降维、聚类和分类模型训练。

步骤1. 文本表示

对于文本表示，我们使用无监督 SimCSE [7] 预训练语言模型。直接使用从 BERT 获得的句子嵌入执行语义相似度相关任务，效果不佳，这是由于句子嵌入具有各向异性。BERT-flow 通过归一化流将句子嵌入分布映射为各向同性高斯分布；BERT-whitening 则将句子嵌入转换为符合标准正态分布的形式，但无法处理特征之间的非线性依赖关系。基于对比学习的理念，通过拉近相似样本、推远不相似样本，SimCSE 大幅提升了语义文本相似度任务中的先进句子嵌入效果。因此，我们直接采用无监督 SimCSE 获取句子嵌入。

步骤2. 降维

经过 SimCSE 嵌入后，句子被表示为 768 维向量。由于大多数聚类算法无法很好地处理高维向量，而且高维向量通常需要较高的计算成本，因此有必要在聚类前对文本嵌入进行降维。降维不仅有助于数据处理，还能帮助可视化文本嵌入的分布。在这一步中，我们采用一种基于流形的降维算法——UMAP。该算法基于黎曼几何和代数拓扑开发，具有坚实的数学基础。与 t-SNE 相比，UMAP 不仅能保留更多全局结构，运行速度也更快。UMAP 中最常用的两个参数是邻居数和最小距离，调整这两个参数可以有效控制降维结果中局部

结构与全局结构之间的平衡。邻居数参数通过限制局部邻域的大小，控制 UMAP 如何平衡局部结构与全局结构。较小的值会使 UMAP 更注重局部结构，而较大的值则会使 UMAP 更倾向于呈现全局结构，同时舍弃细节。最小距离参数表示低维空间中点与点之间的最小距离，该参数控制 UMAP 允许点聚集在一起的紧密程度。取值较低时，低维嵌入空间中的点会呈现更精细的拓扑结构；取值较高时，UMAP 会使点分布得更松散，并更注重保留整体拓扑结构。

步骤3. 聚类

由于低维空间中的数据分布形状不规则、密度分布不均匀，且存在噪声点，我们采用基于密度的聚类算法 HDBSCAN。HDBSCAN 并非寻找具有特定形状的簇，而是寻找比周围空间更稠密的区域；它能够处理噪声，即使簇的形状任意，也能将其找出。HDBSCAN 有两个重要参数：度量和最小簇大小。度量参数用于衡量向量之间的距离，本文所有实验均使用欧几里得度量。最小簇大小参数表示簇的最小大小，也会影响主题数量。该值越大，簇的数量越少，反之亦然。在我们的实验中，会根据不同数据集的特点调整最小簇大小。

与主框架中的降维和聚类步骤可以采用任意的降维与聚类算法组合不同，为了保证伪标注数据的高质量，我们在此步骤中特别选择将 UMAP 与 HDBSCAN 结合使用。这种组合利用了 UMAP 保留数据全局结构和拓扑结构的能力；此外，由于 HDBSCAN 能够对不同密度的数据进行聚类，因此可以自然地选取高密度区域中的非噪声点作为伪标注数据。文本嵌入经过降维和聚类后，每个非噪声数据点都会获得标注，并可用于训练分类模型。

步骤 4：分类模型训练

簇特征增强语言模型可以通过训练分类模型获得，该分类模型由原始预训练语言模型和顶部附加的全连接分类器组成。增强语言模型可以先通过伪标注数据训练分类模型，再移除训练好的分类模型中的全连接

层。与原始预训练语言模型相比，增强语言模型可以通过分类任务的训练获得更多簇语义信息，从而能够生成簇结构更加明显的句子嵌入。

基于增强语言模型的文本聚类

由于增强语言模型具有更强的簇语义表征能力，因此更有利于降维和聚类，对算法选择的要求也较低。我们主要框架中的降维和聚类算法并不局限于 UMAP 与 HDBSCAN 的组合，也可以替换为不同的降维和聚类算法。事实上，框架这一部分的关键在于如何获得具有更明显簇结构的语言模型。借助增强语言模型的强大能力，得到的文本嵌入能够更好地适应后续降维和聚类算法的不同组合。在实验部分，我们将比较不同降维和聚类算法组合的效果。

3.2 基于注意力机制的主题提取

基于增强语言模型进行降维和聚类能够取得良好的文本聚类结果，这表明从增强语言模型中获得的嵌入已包含主题语义信息，而这些主题语义体现在语言模型的某些注意力层中。具体而言，某些注意力层能够更多地关注与主题语义相关的词元，并赋予其更高的注意力权重。这也符合直观假设：某个词越能代表句子的含义，句中其他词对该词的注意力权重就越高。受此启发，我们提出一种基于注意力机制的主题提取方法，利用语言模型的中间产物——每个词元的注意力权重，来衡量该词元在句子中的重要性。对于给定的句子，我们的方法首先对 Transformer 第10层中所有词元指向给定词元的注意力权重求和，作为该词元的重要性评分；然后选取评分最高的词元作为该句子的关键词；最后，对于给定的簇，按照关键词的频率对所有关键词排序，并选取排名前-K 个作为主题词。具体步骤如下。

步骤1。句子钥匙词元提取

对于给定的句子，我们假设该句子中各词元的值为 $\{\mathbf{v}_1, \mathbf{v}2, ..., \mathbf{v}_n\}$,

矩阵 $\{\alpha_{i,j}\}_{i,j=1}^{n}$ 表示由自注意力层得到的注意力权重矩阵，其中 $\alpha_{i,j}$ 表示从第 i 个词元到第 j 个词元的注意力分数。通过加权求和，我们可以得到第 i 个词元的表示，如下所示：

$$\mathbf{e}_i = \sum_{j=1}^{n} \alpha_{i,j}\mathbf{v}_j, \quad i = 1, 2, ..., n.$$

句子的向量表示可以通过对上述获得的词向量取平均值来得到：

$$\mathbf{e} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{i,j}\mathbf{v}_j = \frac{1}{n}\sum_{j=1}^{n}(\sum_{i=1}^{n}\alpha_{i,j})\mathbf{v}_j = \frac{1}{n}\sum_{j=1}^{n}\beta_j\mathbf{v}_j,$$

其中 $\beta_j = \sum_{i=1}^{n}\alpha_{i,j}$ 表示句子中所有词元指向第 j 个词元的注意力权重之和。直观地说，如果某个词元能够充当句子的钥匙，它就会受到句中所有词元更多的注意。换言之，句子中的所有词元都会对这个钥匙词元给予较多注意，因此，根据钥匙词元（第 j个词元）得到的 β_j 值会较大。需要注意的是，上述分析在数学上并不严谨，仅用于为我们的直觉提供一些分析上的支持。

我们还可以通过注意力图将上述分析可视化，以了解词元之间的关联性，以及每个词元在句子中的重要性。图2展示了句子“多家银行封杀信用卡支付宝交易（许多银行已禁止在支付宝使用信用卡进行交易）”的注意力图。线条的粗细表示从左侧词元指向右侧词元的注意力权重，线条越粗，注意力权重越大。根据上述定性分析的结果，Transformer 的第10层已经学到了簇语义，并且能够对与簇主题相关的词元给予更多注意。然而，上述结果仅展示了单个句子的可视化，并不能代表整体情况。因此，我们将在下文中进行统计分析，以比较不同 Transformer 层对簇语义的敏感程度。

表1展示了基于增强语言模型的 Transformer 各层中注意力权重的多种统计指标。我们选取变异系数（Cov.）、峰度（Kurt.）和相对极差（R.R.）作为统计指标，从不同角度反映数据是否存在焦点。这样做的目的是选出承载更多簇语义的合适 Transformer 层，从而为每个句子提供关键词。变异系数是衡量数据相对变异程度的指标，计算方法是用标准差除以均值。数值越小，表示数据点越接近均值；数值越大，则表示数据点越分散。峰度用于衡量分布的尖峭程度。

它衡量数据相对于分布尾部集中在均值附近的程度，并揭示数据集中极端值的相对水平。峰度值越高，概率分布越尖峭；峰度值越低，分布越平坦。极差是数据中最大值与最小值之差。相对极差是极差与数据均值之比，可用于表示离散程度。

层	0	1	2	3	4	5	6	7	8	9	10	11
Cov.	0.1767	0.2422	0.2723	0.3001	0.5308	0.4726	0.4228	0.6646	0.7258	0.8766	1.2354	0.2855
峰度	0.4379	0.5333	0.6412	0.4578	2.7852	0.1905	1.8723	3.9464	4.2924	5.6261	8.2876	0.6091
R.R.	0.6737	0.9415	1.0570	1.1436	2.1683	1.9643	1.6809	2.7304	2.9495	3.6941	5.4171	1.0846

表1：不同层中注意力权重总和的统计指标。Cov.：所有句子的变异系数平均值；Kurt.：所有句子的峰度平均值；R.R.：所有句子的相对极差平均值。每项统计指标中的最高值均已加下划线。

表1中的结果表明，Transformer 第10层对应的注意力权重总和比其他层具有更高的离散程度、峰态和极差，这说明数据中存在明显的峰值，而这些峰值对应句子的关键词。与其他层相比，Transformer 第10层对应的注意力权重具有更高的 Cov.、Kurt. 和 R.R.，这说明数据中存在明显的峰值，而这些峰值对应句子的关键词元。基于上述定性和定量分析，我们选取 Transformer 第10层的注意力权重来确定句子的关键词。

步骤2。句子关键词提取

在此步骤中，我们将步骤1获得的关键词元映射到其对应的词语，以获取每个句子的关键词。首先，我们使用 THULAC 进行句子分词。然后，基于分词后的句子，将步骤1获得的关键词元映射到其对应的词语，得到关键词列表。例如，对于句子“楼 上 下水 管 道 漏 水 , 滴 到 楼 下 家 里（楼上的水管漏水，滴到了楼下的家里） ”，步骤 1 中评分最高的词元是 “ 漏（ leaking）”，分词后的句子为“楼上/下水/管 道/漏水/,/滴/到/楼下/家里/”，将“漏”映射为“漏水”，最终得到该句子的关键词。

步骤3。聚类主题词提取

在此步骤中，我们将提取每个簇的主题词。对于给定的簇，首先汇总所有句子的关键词，构建该簇的关键词集合。然后，按照词频从高到低对关键词进行排序。最后，取列表中的前 K 个关键词作为该簇的主题词。

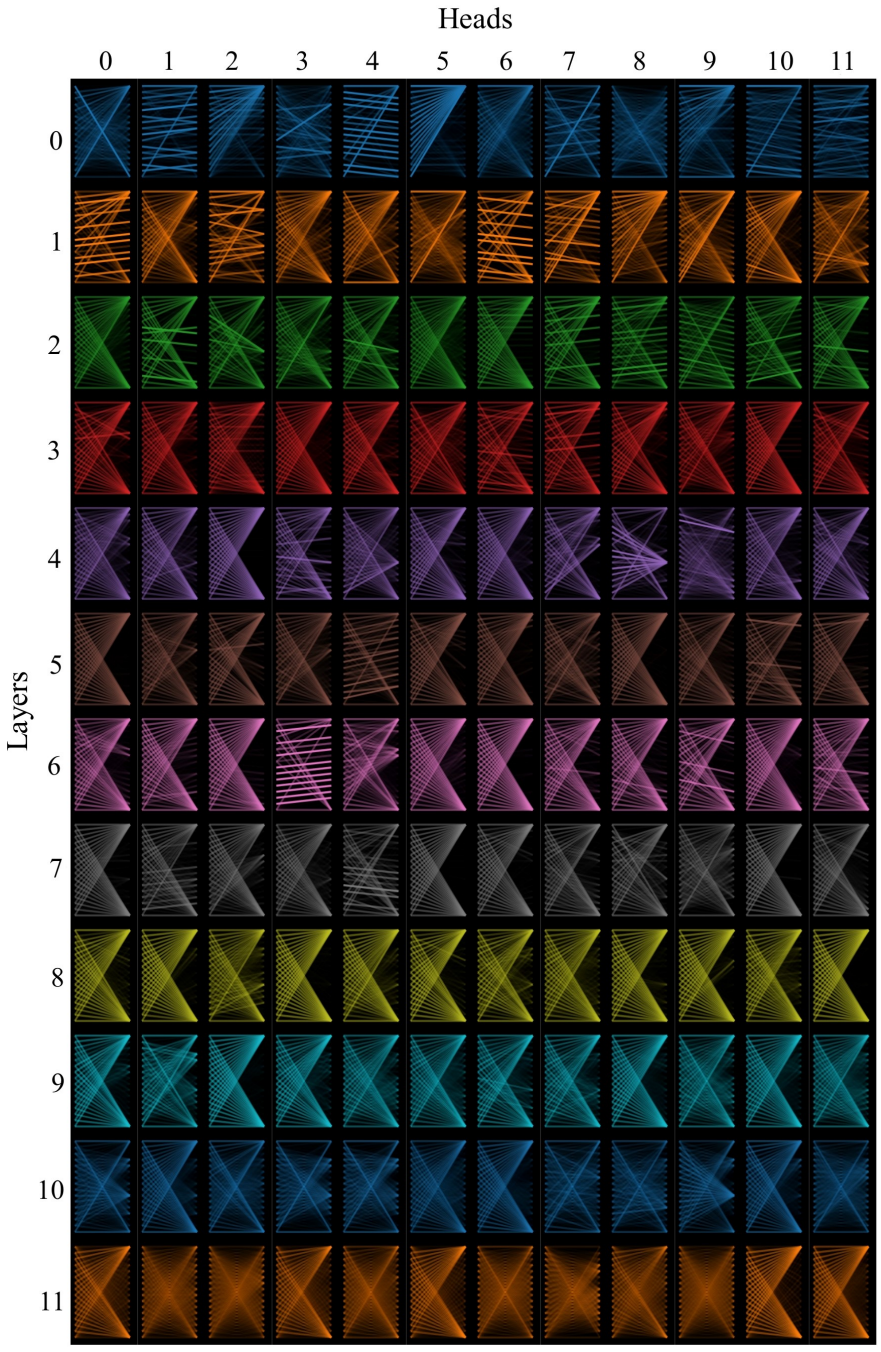


图2：该图展示了 enhanced-SimCSE-chinese-roBERTa 模型对句子 '多家银行封杀信用卡支付宝交易（多家银行已禁止通过支付宝使用信用卡进行交易）' 所生成的 12 层 Transformer 注意力权重，其中列表示 Transformer 层，行表示注意力头。

4 实验

4.1 实验设置

数据集

我们使用两个数据集验证所提出的框架：THUCNews数据集 [29]由清华大学自然语言处理实验室提供，另一个是由当地社区网格员收集的社会治理数据集。THUCNews数据集由筛选2005年至2011年新浪新闻 RSS 订阅频道的历史数据生成，包含74万篇新闻文档（2.19 GB），分为14个类别：金融、彩票、房地产、股票、家居、教育、科技、社会、时尚、政治、体育、星座、游戏和娱乐。我们从每个类别中选取1000条标题数据，组成一个共计14000条数据的数据集。社会治理数据集包含社区网格员收集的有关生活环境、交通、公共设施和邻里纠纷的社会治理事件，共计16870条。

参数选择

实验过程中，我们在调整 UMAP 和 t-SNE 中的邻居数、DBSCAN 和 HDBSCAN 中的最小簇大小，以及 DBSCAN 中的邻域半径时，综合考虑数据集规模和类别数量。本文在调整 UMAP 和 t-SNE 中的邻居数、DBSCAN 和 HDBSCAN 中的最小簇大小，以及 DBSCAN 中的邻域半径时，同时考虑了数据集规模和类别数量。我们在 THUCNews数据集上遍历邻居数和最小簇大小参数，取值范围为100–200、步长为10；遍历邻域半径参数，取值范围为0–4、步长为0.1，并交叉组合这些参数，选取标准化互信息最高的参数组合。考虑到社会治理数据集主题的复杂性和多样性，我们设置最小簇大小=20 和邻居数=20，以确保算法能够提取更细粒度的类别信息。

4.2 文本聚类

实验中用于比较的基线模型包括 LDA [24], FastText [30] 和 Top2Vec[26]。基于所提出的框架，我们选择 Bert-base-Chinese[1] （BERT）、SimCSE[2],及其增强版本（增强版 BERT 和增强版 SimCSE）作为文本嵌入模型；选择 UMAP、t-SNE 和主成分分析作为降维算法；选择 K-Means、DBSCAN 和 HDBSCAN 作为聚类算法，并比较不同模型与算法交叉组合的效果。为了全面评估

[1]https://huggingface.co/bert-base-chinese/tree/main[2]
https://huggingface.co/cyclone/SimCSE-chinese-roberta-wwm-ext/tree/main

所提出框架的性能，我们采用以下外部评价指标：调整兰德指数（ARI） [31],调整互信息（AMI） [32], 、纯度和 F1度量；内部评价指标包括：轮廓系数（SC）、Calinski-Harabasz指数（CHI）和戴维斯-博尔丁指数（DBI）。

表2展示了不同方法在THUCNews数据集上的聚类结果。我们选择LDA、FastText和Top2Vec作为基线，并比较语言模型（即原始语言模型和增强语言模型）、降维算法与聚类算法的不同组合。与基线模型相比，基于语言模型的不同组合在各项评估指标上表现更好。在使用相同降维算法和聚类算法的情况下，增强语言模型能够取得比原始语言模型更好的聚类结果。例如，与BERT+PCA+HDBSCAN相比，BERT- enhanced+PCA+HDBSCAN的调整兰德指数绝对提升了 +26.68%，这进一步验证了增强语言模型生成的文本嵌入具有更强的聚类特性。使用t-SNE或UMAP进行降维时，增强语言模型的表现并未显著优于原始语言模型，这是因为t-SNE和UMAP能够保留数据的全局和局部结构。因此，即使原始语言模型生成的文本嵌入没有明显的聚类特性，由于空间分布特征得以保留，仍然可以获得可接受的聚类结果。

表2还显示，在语言模型增强之前，BERT的表现优于SimCSE。例如，BERT+t-SNE+K均值和BERT+UMAP+HDBSCAN在多项外部评价指标上取得了最佳结果，而BERT+UMAP+HDBSCAN也是BERTopic使用的组合。BERT+UMAP+K均值在多项内部评价指标上表现最佳，且略优于SimCSE+UMAP+K均值。语言模型增强后，增强版 BERT+(t-SNE/UMAP)+(K均值/DBSCAN/HDBSCAN)在外部评价指标上的表现更好，而增强版 SimCSE+UMAP+HDBSCAN在内部评价指标上取得了最佳表现。

图3展示了在固定 UMAP 与 HDBSCAN 组合算法的情况下，增强 BERT（或 SimCSE）前后得到的文本表示的聚类结果。可以看出，与原始语言模型相比，增强语言模型得到的嵌入空间中，簇的边界更加明显，不同簇之间的距离更远，同一簇中的点也更加紧密。此外，增强 SimCSE 得到的簇在太空中比增强 BERT 得到的簇分布更均匀，形状也更接近球体，这进一步验证了表2中的结论：SimCSE 的内部评价指标

优于 BERT。

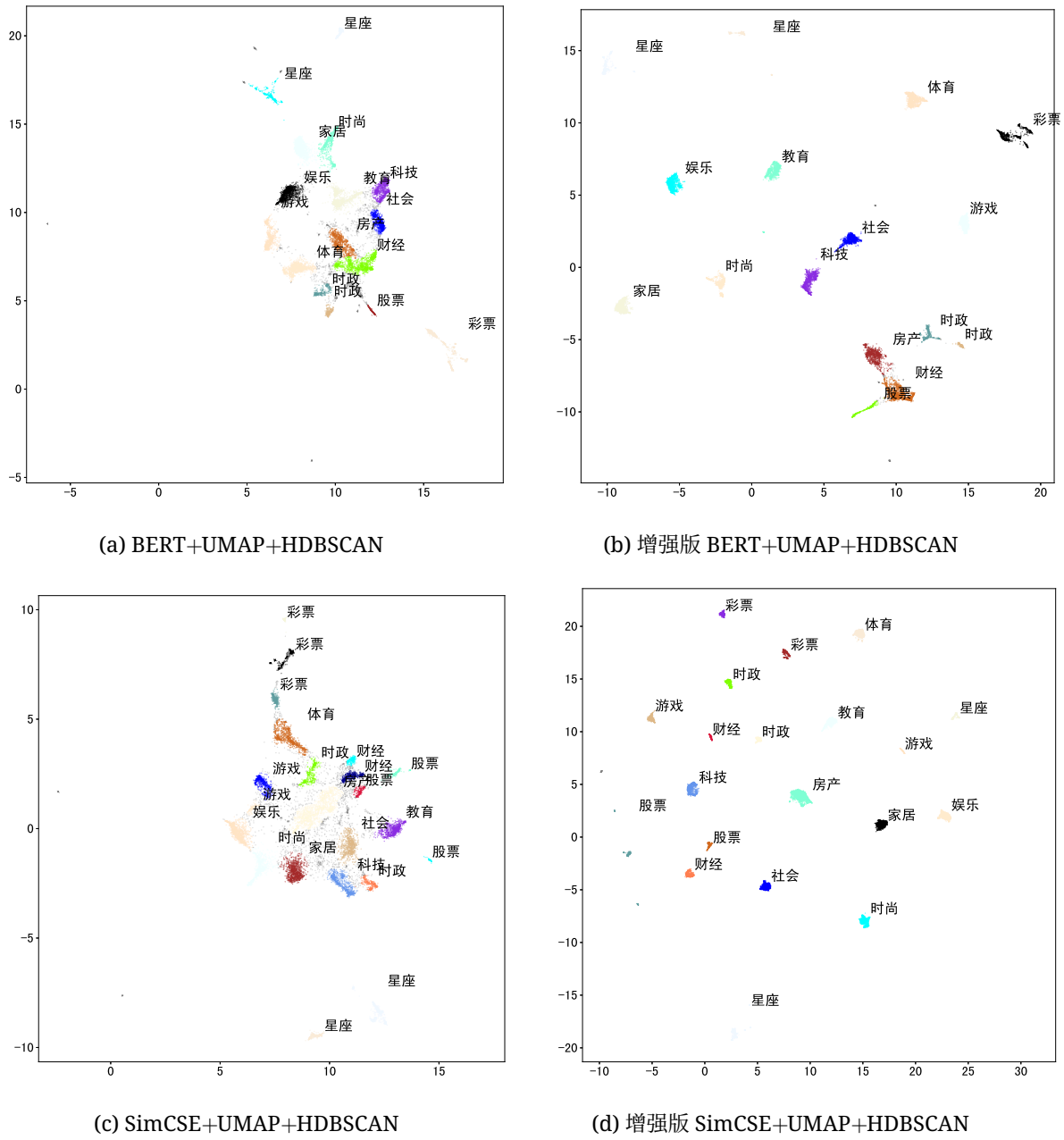
由于 THUCNews 的真实标签只能反映数据集的一种可能划分方式，不能排除存在更优划分方式的可能性，并且考虑到不同聚类数组合的影响，BERT 在所有外部评价指标上均优于 SimCSE，并不一定意味着其实际聚类能力更强。在现实场景中，由于数据通常没有真实标签，因此在将数据划分为类别时，我们更多地会考虑数据本身的分布特征，并希望获得如图3所示的增强SimCSE的嵌入结果，使其在各个类别中均匀分布。因此，对于现实场景中的数据，我们选择内部评价指标更优的 SimCSE 进行文本嵌入。表3展示了 SimCSE 和增强SimCSE 在社会治理数据集上的聚类结果。可以看出，包含增强语言模型的组合显著提高了召回率和 F1值。这主要是因为增强语言模型能够提升文本语义的簇表示能力，减少原始表示空间中簇之间的边缘点数量，并减少 HDBSCAN 聚类过程中噪声点的生成。社会治理数据集通过人工判断聚类结果的合理性进行标注。

4.3 主题提取

表4展示了所提出的基于注意力的方法和 BERTopic 中的 c-TF-IDF算法在社会治理数据集和 THUCNews数据集上学习到的主题，其中注意力机制来自 BERT、SimCSE 及其增强版本。为突出不同主题提取方法之间的比较，本文分别从社会治理数据集和 THUCNews数据集中选取5类标注数据，组成两个具有真实标签的子数据集。表4中的增强版 BERT和 SimCSE，是在这些子数据集上训练分类模型20轮后得到的。从表4可以看出，与 c-TF-IDF相比，基于增强语言模型的注意力机制方法提取的前5个关键词更能体现类别的核心语义。例如，c-TF-IDF在社会治理数据集上提取了大量无意义的词语，如“小区（小区）”“居民（居民）”“社区（社区）”和“情况（情况）”等，而 SimCSE-enhanced+Attention 提取的词语都与主题相关。尽管 BERT-enhanced+Attention 也提取了一些与类别主题无关的词语，如“影响（影响）”和“双方（双方）”，但其整体效果仍优于 c-TF-IDF。

模型	ARI	AMI	纯度	F1	SC	CHI	DBI
LDA	0.021712650	0.043507083	0.156500000	0.094070374	-	-	-
FastText	0.036345435	0.279554411	0.476142857	0.116871716	0.165747600	263.7635306	1.325474649
Top2Vec	0.053700707	0.189647257	0.306285714	0.139114747	0.059875946	1358.713134	1.078640084
BERT+主成分分析+K均值	0.112950249	0.263254817	0.303714286	0.179526423	0.333537489	10210.14964	0.811310507
BERT+主成分分析+DBSCAN	0.030189282	0.126263402	0.174285714	0.147755171	-0.514981608	41.01350480	1.459738375
BERT+主成分分析+HDBSCAN	0.003900224	0.116886870	0.188928571	0.123785626	-0.447570568	36.22244089	1.502777369
BERT+t-SNE+K均值	**0.658144084**	**0.732398618**	**0.807000000**	**0.682909235**	0.484918600	19709.90657	0.698602062
BERT+t-SNE+DBSCAN	0.562179536	0.711425242	0.721857143	0.595534269	0.312350720	6251.544325	1.871509697
BERT+t-SNE+HDBSCAN	0.522267486	0.697662961	0.665357143	0.560068498	0.298417870	5731.753357	1.553215702
BERT+UMAP+K均值	0.594205812	0.723575557	0.706785714	0.627605676	**0.589915900**	**31211.98794**	**0.472229917**
BERT+UMAP+DBSCAN	0.571411301	0.711886810	0.780357142	0.602886097	0.5070201000	2681.931503	1.744340103
BERT+UMAP+HDBSCAN[2]	**0.642479532**	**0.740278657**	**0.837714286**	**0.667820329**	0.510587200	4447.298455	1.722989958
SimCSE+主成分分析+K均值	0.106347177	0.252775170	0.288357143	0.175183474	0.344200357	11244.445811	0.815459048
SimCSE+主成分分析+DBSCAN	0.044355821	0.151427560	0.159000000	0.164180209	-0.254810313	230.1634931	2.394114761
SimCSE+主成分分析+HDBSCAN	0.002175732	0.113755726	0.192142857	0.123710760	-0.424716346	35.80529592	1.956265486
SimCSE+t-SNE+K均值	0.547095723	0.661444030	0.729714286	0.581130327	0.474737140	18324.75306	0.645366147
SimCSE+t-SNE+DBSCAN	0.531404366	0.679330645	0.784571429	0.562565842	0.360822170	5695.250806	1.470280551
SimCSE+t-SNE+HDBSCAN	0.380231096	0.653010575	0.726071429	0.429845375	0.282251420	3646.129072	1.513032985
SimCSE+UMAP+K均值	0.498917295	0.650605537	0.659714286	0.540381135	**0.540319100**	**24314.62983**	**0.491694285**
SimCSE+UMAP+DBSCAN	0.552333818	0.679338022	0.792357143	0.582409120	0.521376670	5455.263584	1.182661660
SimCSE+UMAP+HDBSCAN	0.509769071	0.674170452	0.790928571	0.542680251	0.480586700	5832.627975	1.805907217
增强版 BERT+主成分分析+K均值	0.446185093	0.577556915	0.597642857	0.489695836	0.513355760	28930.99599	0.703627083
增强版 BERT+主成分分析+DBSCAN	0.306394174	0.528642067	0.452000000	0.374108656	0.388692892	6922.794374	1.276219364
增强版 BERT+主成分分析+HDBSCAN	0.270715954	0.530203903	0.528785714	0.340734720	0.182775500	3178.215554	1.462026462
增强版 BERT+t-SNE+K均值	0.669056669	0.744554080	0.816142857	0.693527327	0.634305800	35938.15804	0.500305803
增强版 BERT+t-SNE+DBSCAN	**0.713460719**	0.768681769	0.872357142	0.732967082	0.623143850	21405.49619	1.457663276
增强版 BERT+t-SNE+HDBSCAN	**0.719518003**	0.771625801	**0.875285714**	**0.738656874**	0.631086800	26618.08005	0.783802092
增强版 BERT+UMAP+K均值	0.698639960	**0.774216205**	0.830785714	0.721363204	0.808400150	157304.9432	**0.247107253**
增强版 BERT+UMAP+DBSCAN	0.711797510	0.771620123	0.870642857	0.731856021	0.821047100	26382.62055	2.287925582
增强版 BERT+UMAP+HDBSCAN	0.718627893	**0.773906464**	**0.877428571**	**0.738152439**	0.830618300	98233.07291	1.142740436
增强版 SimCSE+主成分分析+K均值	0.391048446	0.501399593	0.564714286	0.438903832	0.550941828	45130.16688	0.659255520
增强版 SimCSE+主成分分析+DBSCAN	0.230694030	0.500170841	0.500571429	0.305328603	0.252246800	3402.093261	1.321582415
增强版 SimCSE+主成分分析+HDBSCAN	0.290785881	0.512756342	0.536500000	0.356713658	0.362360205	11606.72156	2.352973108
增强版 SimCSE+t-SNE+K均值	0.524010214	0.649913615	0.703285714	0.559648674	0.672626400	31379.05478	0.542381072
增强版 SimCSE+t-SNE+DBSCAN	0.586704695	0.705531691	0.842071429	0.614094205	0.785797830	83140.03355	1.260080529
增强版 SimCSE+t-SNE+HDBSCAN	0.590634454	0.704936408	0.843714286	0.617538117	0.760994260	84049.29390	0.401127394
增强版 SimCSE+UMAP+K均值	0.584538557	0.687672859	0.755071429	0.615344342	0.788089900	42994.13445	0.455941761
增强版 SimCSE+UMAP+DBSCAN	0.587115673	0.705317471	0.842214286	0.614454433	**0.919167700**	**285574.0943**	0.378244737
增强版 SimCSE+UMAP+HDBSCAN	0.591524314	0.707746769	0.842214286	0.618695607	**0.912691300**	**269272.5169**	**0.163748030**

表2：THUCNews数据集上的聚类评估。该表展示了三种基准模型及不同模型组合在文本表示、降维和聚类框架下于THUCNews数据集上的表现。所有降维和聚类算法的参数均根据上述参数选择方法采用最优参数。在所有组合中，K均值均选取THUCNews数据集的真实类别数量14作为 k 值。“BERT”表示 BERT-base-chinese模型，“SimCSE”表示 SimCSE-chinese-roBERTa模型，“enhanced”表示增强模型，“BERT+UMAP+HDBSCAN”表示 BERTopic 使用的组合方法。表中分别以粗体标出改进前后排名前两位的指标数据。在各项评估指标中，只有戴维斯-博尔丁指数（DBI）越小，聚类结果越好；其他指标则相反。



(a) BERT+UMAP+HDBSCAN

(b) 增强版 BERT+UMAP+HDBSCAN

(c) SimCSE+UMAP+HDBSCAN

(d) 增强版 SimCSE+UMAP+HDBSCAN

图3：BERT 和 SimCSE 在 THUCNews数据集上增强前后的聚类效果对比。图中中文词语的英文翻译如下：彩票（lottery）、体育（sports）、时政（political science）、财经（finance）、游戏（game）、股票（stock）、房产（real estate）、娱乐（entertainment）、社会（society）、教育（education）、时尚（fashion）、家居（home furnishings）、科技（technology）、星座（constellation）。

模型	精确率	召回率	准确率	F1
SimCSE+UMAP+HDBSCAN	**0.938622883**	0.68359106	0.647243628	0.791059915
增强版 SimCSE+UMAP+HDBSCAN	0.923798212	**0.946822142**	**0.875992887**	**0.935168486**

表3：社会治理数据集上的聚类结果。该表展示了预训练语言模型增强前后的结果。SimCSE 表示 SimCSE-chinese-roBERTa模型，“增强”表示增强模型。每项指标的最大值以粗体标出。

尽管 c-TF-IDF 采用基于类别的 TF-IDF，能够降低同时出现在多个类别中的高频词的重要性排名，但它在以下两个方面存在局限。首先，当类别数量较少时，IDF 的作用会减弱。例如，“组图（图片）”同时出现在“电影”“留学”和“科学”这三个类别中。然而，类别总数只有 5 个，相对较少，因此 IDF 的影响较小，最终仍无法避免高词频带来的影响。其次，如果某个类别中存在一个与类别无关的高频词，且该词没有出现在其他类别中，那么这个高频词仍会对结果产生很大影响，很容易被误认为该类别的中心词。例如，“纠纷”是“土地”类别中的高频词，但它并未大量出现在其他类别中，因此其 IDF 权重很小。最终，该词会主要受到“纠纷”的词频影响，并被视为该类别的关键词。基于注意力的主题提取的优点在于，在获得增强语言模型的过程中（即分类模型训练过程中），模型已经学会了类别之间的差异，因此增强语言模型会将注意力转向与类别相关的词语。例如，增强前的语言模型更容易关注“纠纷”，但由于“纠纷”也出现在不同类别中，这说明“纠纷”是不同类别的共同特征。在增强语言模型的训练过程中，模型对共同特征的注意力会逐渐减弱，最终转向能够区分不同类别的代表性关键词。这避免了提取“纠纷”“居民”和“组图”等无关词语的现象。

在基于注意力机制的四种组合中，增强语言模型版本提取出的主题相关词比例高于原始版本。如表4所示，在社会治理数据集的五个类别上，BERT（或 SimCSE） + 注意力机制提取出了9个（或4个）与类别无关的词，而 BERT（或 SimCSE）-增强版 +注意力机制仅提取出2个（或0个）词。随着语言模型得到增强，关键词的重要性排序也发生了变化。例如，SimCSE 增强后，其对“失物”类别的关注重点从“钥匙”变为与类别主题最相关的“丢失”。此外，表4还表明，语言模型 SimCSE 的平均表现优于 BERT。

由于所提出的基于注意力的主题提取方法和 c-TF-IDF 方法都依赖句子分词，且会受到分词器的影响，因此提取出的关键词可能包含“绳狗屎”和“冠状病毒疫苗”等不规范词语。

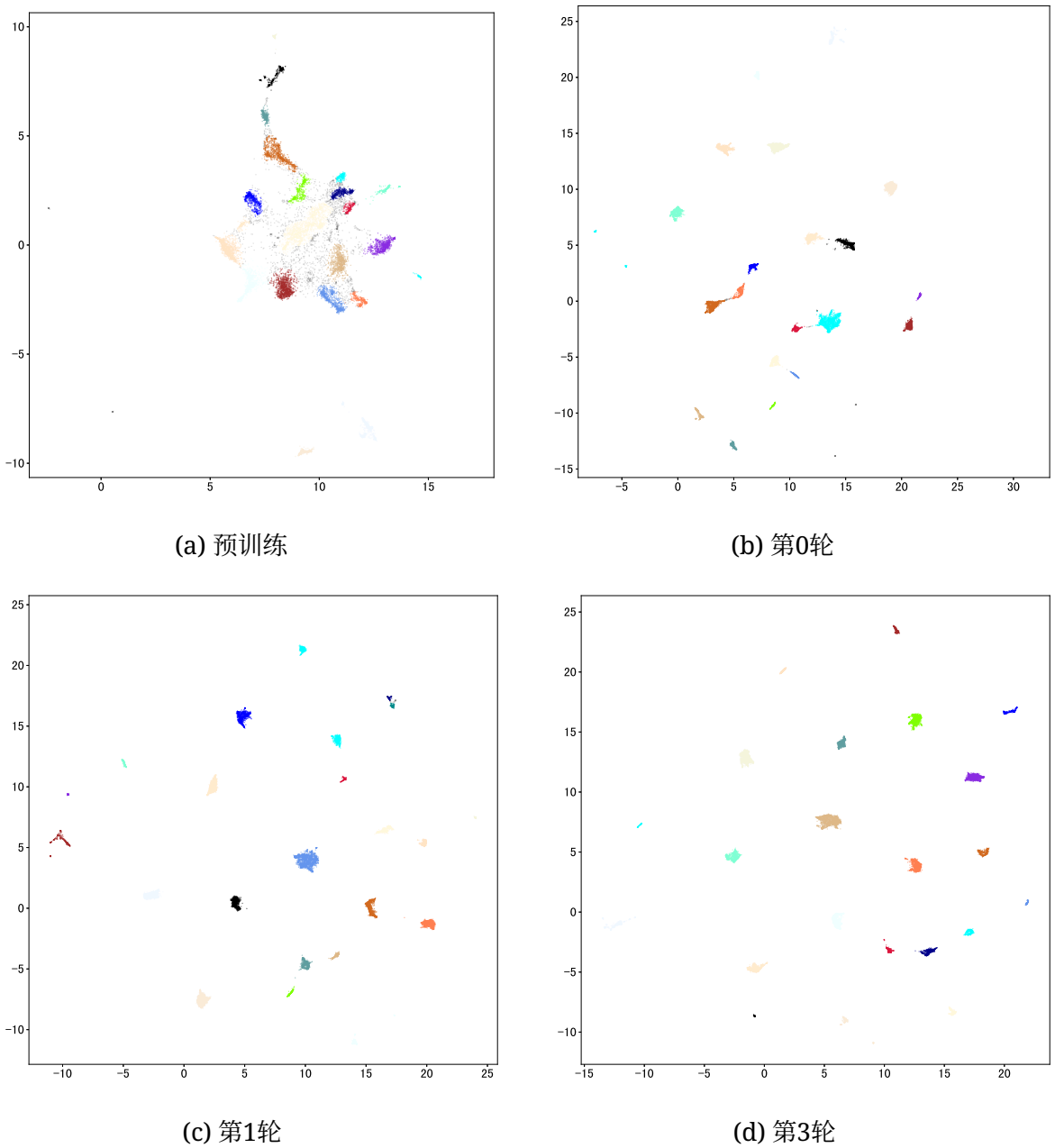
方法	社会治理数据集					THUCNews				
	遛狗	丢失物品	疫苗	噪音	土地	足球	电影	彩票	留学	科学
	(遛狗)	**(失物)**	**(疫苗)**	**(噪声)**	**(土地)**	**(足球)**	**(电影)**	**(彩票)**	**(留学)**	**(科学)**
真实类别 +c-TF-IDF	遛狗	钥匙	接种	休息	纠纷	足协	电影	负彩	留学	科学
	(遛狗)	(钥匙)	(接种)	(休息)	(争议)	(足球协会)	(电影)	（负彩票）	（留学）	(科学)
	小区	丢失	疫苗	影响	发生	女足	票房	推荐	美国	发现
	（小区）	（失去）	(疫苗)	(影响)	(发生)	(女子足球)	(票房)	(推荐)	(美国)	(发现)
	狗绳	社区	冠疫苗	楼上	产生	中国	香港	主场	组图	组图
	(狗绳)	(社区)	(新冠疫苗)	(楼上)	（生成）	（中国）	（香港）	（主场）	（图片）	（图片）
	居民	身份证	咨询	居民	土地	足球	新片	开奖	大学	地球
	(居民)	(身份证)	(询问)	(居民)	(土地)	(足球)	（新电影）	（买彩票）	（大学）	（地球）
	栓绳	失主	情况	扰民	问题	韦迪	组图	火线	移民	英国
	（系绳）	（失主）	(情况)	(滋扰)	问题	(Di Wei)	(图片)	(射击线)	(移居国外)	(英国)
BERT+注意力机制	拴绳	丢失	冠疫苗	扰民	纠纷	足协	组图	双色球	组图	组图
	(系绳)	(失去)	(新冠疫苗)	(滋扰)	(争议)	(足球协会)	(图片)	(双层色球)	(图片)	(图片)
	遛狗	失主	疫苗	影响	地界	女足	票房	彩民	留学	月球
	(遛狗)	(失主)	(疫苗)	(影响)	(边界)	（女子足球）	（票房）	（彩票购买者）	（留学）	（月亮）
	狗绳	身份证	咨询	吵闹	村委会	足球	电影	头奖	移民	2008
	（狗绳）	(身份证)	(询问)	(争吵)	(村委会)	(足球)	(电影)	(头奖)	(移居国外)	(2008年)
	绳子	钥匙	情况	噪音	村民	球员	甄子丹	初盘	澳洲	揭秘
	(绳子)	(钥匙)	(情况)	(噪声)	(村民)	(玩家)	(甄子丹)	(初始游戏)	(澳大利亚)	(揭示)
	绳狗屎	忘记	询问	孩子	争执	足坛	打分	负彩	美国	火星
	(绳子狗屎)	(忘记)	(询问)	(儿童)	(争议)	(足球界)	(评分)	(负彩票)	(美国)	(火星)
增强版 BERT+注意力机制	遛狗	钥匙	疫苗	影响	土地	女足	梅兰芳	推荐	留学	科学
	（遛狗）	（钥匙）	（疫苗）	（影响）	（土地）	（女子足球）	(《梅兰芳》)	(推荐)	(留学)	(科学)
	狗绳	丢失	接种	噪音	宅基地	足协	画皮	火线	美国	发现
	(狗绳)	(失去)	(接种)	(噪声)	(宅地)	(足球协会)	(画皮)	(射击线)	(美国)	(发现)
	狗链	身份证	冠疫苗	声音	双方	足球	冯小刚	双色球	留学生	火星
	(狗绳)	(身份证)	(新冠疫苗)	(声音)	(双方)	(足球)	(冯小刚)	(双层色球)	(留学生)	(火星)
	绳子	失主	打疫苗	吵闹	地界	世界杯	叶问	负彩	英国	研究
	（绳子）	（失主）	（接种疫苗）	（争吵）	（边界）	（世界杯）	(叶问)	(负彩票)	(英国)	(学习)
	绳狗屎	钱包	针疫苗	晚上	占地	中国	海角七号	彩票	出国	太空
	(绳子狗屎)	(钱包)	（第n剂疫苗）	（夜晚）	（土地侵占）	（中国）	（海角七号）	（彩票）	(出国)	(太空)
SimCSE+注意力机制	遛狗	钥匙	疫苗	扰民	纠纷	女足	票房	双色球	留学	火星
	(遛狗)	(钥匙)	(疫苗)	(滋扰)	(争议)	(女子足球)	(票房)	(双层色球)	(留学)	(火星)
	拴绳	丢失	冠疫苗	噪音	宅基地	足协	梅兰芳	彩票	移民	机器人
	(系绳)	(失去)	(新冠疫苗)	(噪声)	(宅地)	(足球协会)	(《梅兰芳》)	(彩票)	(移居国外)	(机器人)
	牵绳	失主	咨询	影响	占地	足球	赤壁	米兰	签证	黑洞
	（牵引绳）	(失主)	询问	(影响)	(土地侵占)	(足球)	(赤壁之战)	(国际米兰)	（签证）	（黑洞）
	狗绳	书包	接种	邹居	耕地	世界杯	刘烨	排行榜	申请	月球
	（狗绳）	（书包）	（接种）	(邹居)	(农田)	(世界杯)	(刘烨)	(排名列表)	(申请)	(月亮)
	绳子	被盗	打疫苗	吵闹	地界	主帅	谢霆锋	(国米)	名校	望远镜
	(绳子)	(被盗)	(接种疫苗)	(争吵)	(边界)	(经理)	(谢霆锋)	(国际米兰)	(名牌大学)	(望远镜)
增强版 SimCSE+注意力机制	遛狗	丢失	疫苗	休息	土地	女足	梅兰芳	彩票	留学	科学
	(遛狗)	(失去)	(疫苗)	(休息)	(土地)	(女子足球)	(《梅兰芳》)	(彩票)	(留学)	(科学)
	狗绳	钥匙	冠疫苗	楼上	宅基地	足协	票房	负彩	美国	发现
	(狗绳)	(钥匙)	（新冠疫苗）	(楼上)	(宅地)	(足球协会)	（票房）	（负彩票）	(美国)	(发现)
	拴绳	身份证	接种	扰民	地界	世界杯	香港	推荐	出国	研究
	(系绳)	(身份证)	(接种)	(滋扰)	(边界)	(世界杯)	(香港)	(推荐)	(出国)	(学习)
	绳子	失主	打疫苗	噪音	占地	韦迪	冯小刚	双色球	移民	南极
	(绳子)	(失主)	(接种疫苗)	(噪声)	(土地侵占)	(Di Wei)	(冯小刚)	(双层色球)	(移居国外)	(南极点)
	牵绳	遭失	针疫苗	晚上	耕地	足球	周迅	彩民	留学生	机器人
	(牵引绳)	(失去)	(第n剂疫苗)	(夜晚)	(农田)	(足球)	(Margaret)	(彩票购买者)	(留学生)	(机器人)

表4：不同主题提取算法在社会治理数据集和THUCNews数据集上的效果比较。我们已将各算法提取出的、与类别主题无关的词语加下划线。

4.4 模型增强的效果

簇分布的变化

为探究训练过程中嵌入的空间分布变化，本节使用SimCSE在THUCNews数据集上进行实验，并在图4中展示嵌入空间的变化。



(a) 预训练　(b) 第0轮

(c) 第1轮　(d) 第3轮

图4：模型增强过程中簇的空间分布。该图展示了采用原始预训练SimCSE和增强SimCSE时，THUCNews数据集上二维嵌入的空间分布变化趋势。epoch $i(i = 1, 2, 3)$表示模型训练轮次，标签为簇标签。

这表明，原始预训练语言模型获得的嵌入分布并未呈现出明显的簇——

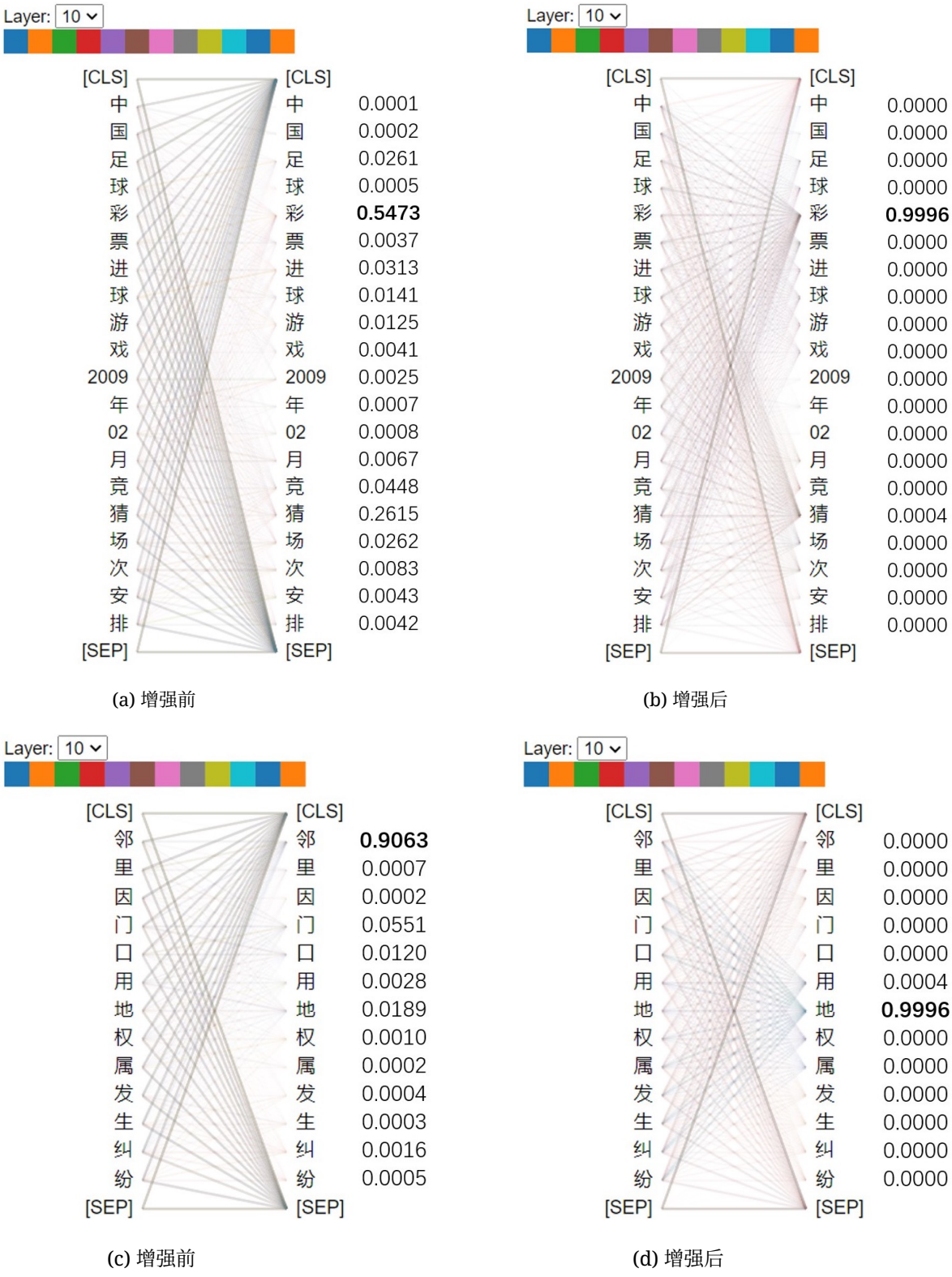
簇形状各异。此外，簇与簇之间没有清晰的边界，且簇形状不规则。随着训练过程的推进，嵌入空间中的簇结构逐渐变得清晰，不同簇的整体分布也变得更加均匀。最终，形成了簇内距离紧凑、簇间距离较大的分布。

注意力机制的转移

如表4所示，与原始预训练语言模型相比，增强模型的注意力机制能够提取出与主题关联更紧密的关键词。本节从不同数据集中选取两个句子，展示增强前后注意力机制的转移情况。图5展示了THUCNews数据集“彩票”类别中的句子“中国足球彩票进 球游戏2009年02月竞猜场次安排（中国足球彩票进球游戏2009年02月竞猜场次安排）”，以及社会治理数据集“土地”类别中的句子“邻里因门口 用地权属发生纠纷（因入口土地所有权产生的邻里纠纷）”的注意力机制转移情况。比较图5(a)和(b)可以看出，增强前后的模型对词元“彩（彩票）”的注意力都最高，但增强模型提高了对类别语义的关注，因此更加聚焦于与类别主题最相关的关键词，注意力权重也从0.5473提高至0.9996。图5(c)和(d)显示，原始预训练模型最关注句子中的词元“邻（邻里）”，但由于模型的训练目标是更好地区分语料库的类别，注意力便更多地转移到了与“土地”类别相关的词元“地（土地）”上。具体而言，词元“地（土地）”的注意力分数从0.0189提高至0.9996，而词元“邻（邻里）”的评分则从0.9063降至0。此外，两组实验还呈现出一个共同点：增强前，模型的注意力较为均匀地分布在句子的各个词元上；增强后，模型几乎完全聚焦于与类别相关的关键词元，这无疑提升了主题提取效果。

5 结论

本文提出了一个无监督文本聚类与主题提取框架 ClusTop。利用语言模型提供的句子嵌入和注意力权重，将聚类与主题提取这两个密切相关的任务紧密结合起来。所提出的框架无需任何人工标注数据或外部知识，并且在预训练语言模型、降维和簇的选择上具有灵活性——



(a) 增强前

(b) 增强后

(c) 增强前

(d) 增强后

图5：该图展示了模型增强前后，Transformer第10层对句子“中国足球彩票进球游戏2009年02月竞猜场次安排”（来自THUCNews数据集）和“邻里因门口用地权属发生纠纷”（来自社会治理数据集）的注意力权重变化。线条颜色越深，表示当前词元从其他词元处获得的注意力权重越高。每张图的右侧显示了每个词元的注意力分数，该分数通过汇总句子中所有词元的注意力权重，然后应用softmax得到。最大值已突出显示。

聚类算法，并具有很强的领域迁移能力。所提出的框架由四个部分组成：基于增强语言模型的文本表示、降维、基于注意力机制的聚类和主题提取，其中，语言模型的增强以及基于注意力的主题提取是我们工作的主要贡献。实验结果表明，在所有聚类评估指标上，增强语言模型的表现均优于原始预训练语言模型。基于注意力的主题提取方法最大限度地发挥了注意力机制的作用，自然地融合了主题提取与聚类，无需之后再叠加额外的主题提取算法。实验结果表明，在与聚类主题语义的一致性方面，我们的基于注意力的主题提取方法优于强大的主题提取基线方法。

由于所提出的框架独立于具体模型，随着语言模型、降维和聚类算法的持续发展，整体聚类和主题提取性能还可以进一步提升。我们通过训练分类模型，提出了一种可行的无监督方法来获得增强语言模型；但从本质上讲，任何能够使语言模型对簇语义更加敏感的方法，都可以用来替代本文中的增强方法。

参考文献

[1] Ming Zhao、Jianli Wang 和 Guanjun Fan。改进文本聚类算法在智能问答系统中的应用研究。载于*2008*年第二届遗传与进化计算国际会议，第 463–466 页。IEEE，2008年。[2] Xuegang Chen、Sheng Duan 和 Luda Wang。互联网舆情聚类分析研究。*Cluster Computing*，22(3):5997–6007，2019年。[3] Andriy Shepitsen、Jonathan Gemmell、Bamshad Mobasher 和 Robin Burke。利用层次聚类实现社交标签系统中的个性化推荐。载于*2008*年*ACM*推荐系统会议论文集，第 259–266 页，2008年。[4]Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018年。

[5] Bohan Li、Hao Zhou、Junxian He、Mingxuan Wang、Yiming Yang 和 Lei Li。关于预训练语言模型生成的句子嵌入。自然语言处理实证方法，第 9119–9130 页，2020年。

[6] Jianlin Su、Jiarun Cao、Weijie Liu 和 Yangyiwen Ou。白化句子表示，以获得更好的语义效果和更快的检索速度。*arXiv* 预印本 *arXiv:2103.15316*，2020年。[7] Tianyu Gao、Xingcheng Yao 和 Danqi Chen。SimCSE：简单的句子嵌入对比学习。*arXiv* 预印本 *arXiv:2104.08821*，2021年。[8] Yuanmeng Yan、Rumei Li、Sirui Wang、Fuzheng Zhang、Wei Wu 和 Weiran Xu。ConSERT：用于自监督句子表示迁移的对比框架。*arXiv* 预印本 *arXiv:2105.11741*，2021年。

[9] Xing Wu、Chaochen Gao、Liangjun Zang、Jizhong Han、Zhongyuan Wang 和 Songlin Hu。ESimCSE：用于无监督句子嵌入对比学习的增强样本构建方法。*arXiv*预印本 *arXiv:2109.04380*，2022年。

[10] Yung-Sung Chuang、Rumen Dangovski、Hongyin Luo、Yang Zhang、Shiyu Chang、Marin Solja˘ci´c、Shang-Wen Li、Wen tau Yih、Yoon Kim 和 James Glass。Diffcse：用于句子嵌入的差异式对比学习。*arXiv*预印本 *arXiv:2204.10298*，2022年。

[11] Leland McInnes、John Healy 和 James Melville。UMAP：用于降维的统一流形逼近与投影。*arXiv*预印本 *arXiv:1802.03426*，2018年。

[12] Leland McInnes、John Healy 和 Steve Astels。hdbscan：基于层次密度的聚类。《开放源代码软件期刊》，2(11):205，2017年。

[13] Ian T Jolliffe 和 Jorge Cadima。主成分分析：综述与最新进展。《英国皇家学会哲学汇刊A辑：数学、物理与工程科学》，374(2065):20150202，2016年。

[14] Badrul Sarwar、George Karypis、Joseph Konstan 和 John Riedl。降维在推荐系统中的应用——案例研究。技术报告，明尼苏达大学明尼阿波利斯分校计算机科学系，2000年。[15] Alan Julian Izenman。线性判别分析。载于《现代多元统计技术》，第237–280页。施普林格，2013年。[16] Bernhard Sch¨olkopf、Alexander Smola 和 Klaus-Robert M¨uller。核主成分分析。载于国际人工神经网络会议，第583–588页。施普林格，1997年。[17] Francis R Bach 和 Michael I Jordan。核独立成分分析。《机器学习研究杂志》，3（7月）：1–48，2002年。

[18] Gaston Baudat 和 Fatiha Anouar。基于核方法的广义判别分析。神经计算，12（10）：2385–2404，2000年。[19] Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。《机器学习研究杂志》，9（11），2008年。[20] Martin Ester、Hans-Peter Kriegel、J¨org Sander、Xiaowei Xu 等。用于在含噪声的大型空间数据库中发现簇的基于密度的算法。载于*KDD*，第96卷，第226–231页，1996年。[21] Juan Ramos 等。使用词频-逆文档频率确定文档查询中的词语相关性。第一届机器学习教学会议论文集，242（1）：29–48，2003年。[22] Bruno Trstenjak、Sasa Mikac 和 Dzenana Donko。采用基于词频-逆文档频率框架的 K 近邻文本分类。*Procedia Engineering*，69：1356–1364，2014年。[23] Rada Mihalcea 和 Paul Tarau。TextRank：为文本理清脉络。*2004*年自然语言处理实证方法会议论文集，第404–411页，2004年。[24] David M Blei、Andrew Y Ng 和 Michael I Jordan。潜在狄利克雷分配。《机器学习研究杂志》，3（1月）：993–1022，2003年。[25] P. Sharma 和 Y. Li。通过自标注进行自监督上下文关键词与关键短语检索。预印本，2019年。[26]Dimo Angelov。Top2Vec：主题的分布式表示。*arXiv* 预印本 *arXiv:2008.09470*，2020年。[27] Maarten Grootendorst。BERTopic：采用基于类别的词频-逆文档频率流程的神经主题建模。*arXiv* 预印本 *arXiv:2203.05794*，2022年。[28] Quoc Le 和 Tomas Mikolov。句子和文档的分布式表示。载于国际机器学习会议，第1188–1196页。PMLR，2014年。[29] M Sun、J Li、Z Guo、Z Yu、Y Zheng、X Si 和 Z Liu。THUCTC：高效的中文文本分类器。*GitHub*代码库，*https://github.com/thunlp/THUCTC*。，2016年。[30] Armand Joulin、Edouard Grave、Piotr Bojanowski 和 Tomas Mikolov。高效文本分类的技巧集。计算语言学协会欧洲分会第*15*届会议论文集，第2卷：427–431，2017年。

[31] Lawrence Hubert 和 Phipps Arabie. 比较划分。《分类学杂志》，2(1):193–218, 1985。

[32] Nguyen Xuan Vinh、Julien Epps 和 James Bailey. 聚类比较的信息论度量：是否需要进行偶然性校正？发表于第*26*届国际机器学习年会论文集，第1073–1080页，2009。