万文桐，黄润才. 基于 Text2Vec_AE_KMeans 的微博话题聚类分析方法[J]. 智能计算机与应用,2025,15(5):82-89. DOI:10.20169/j.issn.2095-2163.250511

基于 Text2Vec_AE_KMeans 的微博话题聚类分析方法

万文桐，黄润才
(上海工程技术大学 电子电气工程学院，上海 201620)

摘 要：传统的话题聚类分析方法使用静态词向量对微博文本进行建模,对微博文本不规范表达、一词多义等特点应对不佳,从而影响聚类效果与话题表述。针对此,提出了一种基于 Text2Vec_AE_KMeans 的深度文本特征提取与聚类的微博话题聚类分析方法。首先,使用基于 MacBert 预训练模型与 CoSENT 文本语句建模方法设计的 Text2Vec 预训练模型,对微博话题文本进行文本语义表示,从而改进静态词向量在文本特征建模方面的不足;然后,通过带有非线性激活函数的 AutoEncoder 降维网络对高维非线性文本特征进行降维;最后,在话题聚类分析的过程中采用 KMeans_C-TF-IDF 算法进行面向微博文本的聚类分析,从聚类簇的角度把握话题分布信息。在真实微博话题数据集上,相较于传统静态词向量建模方法,本文提出的方法在聚类评价指标上表现优异,生成的话题信息可识别性较好。

关键词：话题聚类分析；CoSENT；Text2Vec；自编码器
中图分类号：TP391.1 **文献标志码**：A **文章编号**：2095-2163(2025)05-0082-08

Clustering analysis of Weibo topic based on Text2Vec_AE_KMeans

WAN Wentong, HUANG Runcai
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: Static text feature extraction methods are short in dealing with irregular expressions and polysemy words in Weibo text, and results in poor performance in text clustering and topic representations. This paper gives a topic clustering method based on Text2Vec_AE_KMeans which extracts the deep text features while improving the topic representation result. Firstly, Text2Vec model based on MacBert and CoSENT is used to extract deep dynamic features from Weibo text, which results better in text semantic matching tasks compared to the static text feature extraction methods. Then, AutoEncoder is used to catch the key non-linear features while reducing the compute calculate complexity in clustering afterwards. Finally, for a better performance in clustering and topic representations, a method combines KMeans and C-TF-IDF keyword extraction is used to analyze the clustering result, which reveals the distribution feature in the given Weibo topic. Based on a real Weibo topic dataset, the proposed method shows a better result in clustering metrics and topic representation. The generated topic information is more recognizable.

Key words: topic clustering analysis; CoSENT; Text2Vec; auto encoder

0 引 言

以微博为代表的社交媒体平台在社会热点事件舆情发展与传播的过程中扮演了重要的角色。微博以短文本为载体向微博用户提供了社会热点事件讨论的平台,并通过点赞、评论、转发与微博话题功能加速了事件相关舆情的发展,对及时把握舆情发展提出了挑战。

为了快速把握微博话题中的舆情信息,可以通过文本聚类分析的方法开展研究。为了对微博话题实现文本聚类分析,需要对微博文本进行特征建模与提取。传统的静态文本特征提取方式如 TF-IDF、Word2Vec 以及 Glove 等基于概率模型方式对文本进行分析,构建词向量并获取句、段向量表示。林江豪等学者[1]通过 Word2Vec 算法获取微博的文本特征表示,对新闻话题下微博评论进行聚类分析。Béatrice 等学者[2]通过 Word2Vec、TF-IDF 的方式对法文推特进行建模,并通过聚类的方式获取推文中

作者简介：万文桐(1996—),男,硕士研究生,主要研究方向:自然语言处理,数据挖掘。
通信作者：黄润才(1966—),男,博士,副教授,主要研究方向:计算机网络与信息安全,智能计算,服务机器人,大数据等。Email:hrc@sues.edu.cn。

收稿日期：2023-09-20

的取话题信息。然而，静态文本特征受限于模型及算法的影响，对微博文本中的不规范表达、一词多义等问题难以处理，且受限于语料量的限制，无法得到泛化的文本特征表示，从而对后续分析过程造成影响。

随着深度学习技术在自然语言处理领域的发展，以 Bert[3-5] 为代表的预训练模型被广泛应用于各种自然语言处理任务。Bert 通过 Transformer 结构，实现了获取带有上下文语境的词向量表示。在话题聚类分析领域，Oguzhan[6] 以公众健康推文为研究对象，通过 Bert 预训练模型对推文进行建模，并通过 K-Means 聚类的方式获取推文话题信息。Anwar 等学者[7] 以 2020 年美国大选其间的推文为研究对象，通过构建词云与 Bert 建模的方式进行话题分析。刘梦颖等学者[8] 通过构建频繁词-Bert 的文本双表示模型对微博文本进行特征提取，并通过谱聚类算法进行微博热点话题的发掘分析。此类方法改善了传统静态词向量模型因为语料库的缺乏所导致的泛化能力的不足，但是 Bert 模型的预训练过程中没有对文本语义匹配任务进行优化处理，在聚类过程中仍有损失。

通过对已有研究的分析可知，微博话题聚类分析工作的效果主要依赖于文本特征提取算法能否有效表示话题文本的深度文本特征。本文在已有研究的基础上提出了一种基于 Text2Vec_AE_KMeans 的微博话题聚类分析方法。首先，针对静态词向量的语料泛化能力限制、Bert 词嵌入在预训练与下游任务的不一致性的问题，本文使用基于 MacBert 与 CoSENT 的文本语义表示模型 Text2Vec 对微博博文进行文本语义表示，改进了传统静态词向量在文本特征建模方面的不足。随后，使用带有非线性激活函数的 AutoEncoder 自编码器对高维文本特征进行降维，在保留关键非线性特征的同时减少后续聚类分析所需的计算量。最后，构建了基于 KMeans_C-TF-IDF 的话题聚类分析与关键词表示框架，通过 Kmeans 聚类算法对文本特征快速聚类得到话题簇，并通过 C-TF-IDF 算法进行话题簇关键词的提取，并最终生成可识别理解的话题表示。

1　方法架构

1.1　MacBert

Bert 模型基于大规模语料、MLM（Masked Language Model）掩码语言模型与 Transformer 结构实现了对上下文信息的有效利用，并通过基于字的最小分割实现了词向量的动态提取与深层文本特征提取。基于字的最小分割虽然能够较好地提取文本的动态特征，但是在预测任务中只能预测原词的一部分，从而对模型的性能造成了一定影响。Bert-wwm 模型采用了 WWM（Whole Word Masking）全词掩蔽技术，在训练阶段的分词过程中通过分词器将某个词划分为多个字，而在掩蔽的过程中对这多个字一起掩蔽，从而保留了词结构的完整语义信息。此外，Bert 在训练阶段使用[Mask]对字符进行掩蔽处理，而在字符的预测阶段并未出现[Mask]，这也导致了预训练任务与微调任务的不一致，还使模型性能受到一定影响。

MacBert[9-10] 对上述 2 个问题进行了改进。首先，MacBert 基于 N-gram 模型的理念，使用了 N-gram Masking 的掩蔽策略，通过设置单字（1-gram）到四字短语（4-gram）共 4 种掩蔽策略，按照 40%、30%、20%、10% 的比例对训练文本进行掩蔽，从而实现了短语层面上的掩蔽，在模型的泛化能力与语义信息提取上相较 WWM 取得了更好的效果。此外，对于 Bert 模型中预训练任务与微调任务不一致的问题，MacBert 使用相似词替换 [Mask] 标记，具体而言，对于分词器得到的词-字组合，MacBert 使用 Word2Vec 进行该词的相似度计算，并选取欧氏距离最近的词用作该词的替换，以替换的方式代替掩蔽，保证了上下游任务的一致性。MacBert 在掩蔽过程中所做的改进见表 1。

表 1　Masking 策略改进样本

Table 1　Different Masking strategies sample

语句	Mask 效果
原始语句	对文件或新闻文章总结
Bert Masking	对文 [M] 或新闻文 [M] 总结
WWM	对 [M] [M] 或新闻 [M] [M] 总结
N-gram Masking	对 [M] [M] 或 [M] [M] [M] [M] 总结
Mac Masking	对文档或网站报道总结

1.2　CoSENT

Bert 模型在文本特征提取与下游的文本分类任务中表现较好。对文本聚类任务来说，可以将文本对输入 Bert 并通过平均池化的方式进行文本相似度的计算。然而，随着文本数据集的增大，文本对的数量也会急剧增加，从而导致计算效率的下降，因而 Bert 直接用于语义搜索匹配和无监督聚类任务效果不佳。

针对这一问题，Sentence Bert[11] 采用孪生神经

网络对文本语义匹配任务进行优化。基于孪生神经网络设计的 Bi-Encoder 通过将文本对分别传入 Bert 产生句嵌入进行文本相似度计算,从而在减少计算量的同时实现了对文本相似度计算任务的优化过程。Sentence Bert 的网络结构如图 1 所示。

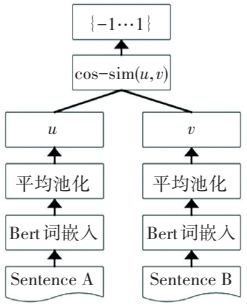

图 1 Sentence Bert 网络结构

Fig. 1 Structure of Sentence Bert network

Sentence Bert, 在计算过程中采用余弦相似度与 *MSE* 均方误差损失函数作为目标函数。余弦相似度与 *MSE* 均方误差损失函数可以表示为:

$$\theta = \cos(u,v) = \frac{u \cdot v}{|u||v|} \tag{1}$$

$$MSE(\hat{\theta}) = E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2 \tag{2}$$

上述的余弦相似度计算式与 *MSE* 目标函数在预训练过程中存在 2 个问题。首先,对于预训练与下游微调任务来说,预训练过程中通过计算余弦相似度的方式构造的损失函数与下游微调过程中进行文本向量表征的目标不一致性使得训练过程的可解释性较差;其次,基于余弦相似度的 *MSE* 目标函数缺少对正负样例的区分,也导致了训练过程的不可控性。

针对 Sentence Bert 的损失函数计算中存在的问题,CoSENT[12] 基于 Circle Loss 多分类损失函数中对交叉熵损失函数的改写,从而将其推广到文本语义匹配任务中。

记 Ω_{pos}、Ω_{neg} 为训练数据中的正负样本对集合,则训练目标可以表示为:

$$cos(u_i,u_j) > cos(u_k,u_l) \tag{3}$$

式(3)的提出旨在区分训练过程中正负样本对对训练过程的影响,从而使得模型能够更好地学习相似样本间的关系。而式(3)通过引入改进的交叉熵计算方式得到了实现。

对单标签分类任务来说,记每个类的得分为 s_1, $s_2,\cdots,s_n$, 目标类为 $t \in \{1,2,\cdots,n\}$, 则交叉熵可以定义为:

$$-\log\frac{e^{s_t}}{\sum_{i=1}^{n} e^{s_i}} = \log\sum_{i=1}^{n} e^{s_i - s_t} = \log\left(1 + \sum_{i=1, i\neq t}^{n} e^{s_i - s_t}\right) \tag{4}$$

在多标签分类场景中,训练目标希望每个目标类得分不小于非目标类的得分,因此可以将式(4)改写为:

$$\log\left(1 + \sum_{i\in\Omega_{neg},\, j\in\Omega_{pos}} e^{s_i - s_j}\right) \tag{5}$$

结合 CoSENT 训练目标式(3),可以导出改进的损失函数如下:

$$\log\left(1 + \sum_{(i,j)\in\Omega_{pos},(k,l)\in\Omega_{neg}} e^{\lambda(\cos(u_k,u_l) - \cos(u_i,u_j))}\right) \tag{6}$$

式(6)通过将分类训练中使用的交叉熵公式推广到以余弦相似度为评判标准的 Sentence Bert 句子相似度任务中,从而有效区分了正负样本对在训练过程中的作用,使得 CoSENT 模型能够更好地学习相似句子对中的特征,从而在文本特征表示上取得更好的效果。

1.3 AutoEncoder 自编码器网络

相较于传统的 Word2Vec 词向量,通过 Text2Vec 预训练模型所得到的文本向量特征维度较高,在进行聚类分析前需要进行降维处理。主成分分析(Principal Component Analysis,PCA)作为常用的特征降维算法,将给定的特征向量通过线性变换转换为另一组特征向量,保留方差大的部分作为降维后的向量表示。其中,基于线性变换的方式使得 PCA 在对深度文本特征等高维非线性特征上的降维效果较差。

随着深度学习技术的不断发展,基于神经网络的降维方法也逐渐被提出。AutoEncoder 自编码器[13]作为一种无监督学习人工神经网络,通过构建编码器-解码器神经网络并嵌入非线性激活函数,使解码器能够最小化从隐藏层的关键特征重构解码器原始输入实现对特征数据维度压缩,从而让隐藏层能够学习原始特征中的非线性特征。一个基础的自编码器结构如图 2 所示。

由图 2 可知,自编码器由编码器(Encoder)以及解码器(Decoder)两部分组成。高维深层文本特征 X 输入编码器后,编码器通过类似于多层感知机的结构将输入的特征进行降维处理,从而得到降维后的深层文本特征 H,而隐藏层的数据作为解码器的输入,解码器再通过与编码器对称的网络结构将降维后的特征 H 进行重构,得到输出的特征重构

$\hat{\boldsymbol{X}}$。这一训练过程可以描述为：

$$H_m = f_{\text{enc}}(W_e X_n + b_e) \tag{7}$$

$$\hat{\boldsymbol{X}} = f_{\text{dec}}(W_h \hat{H} + b_h) \tag{8}$$

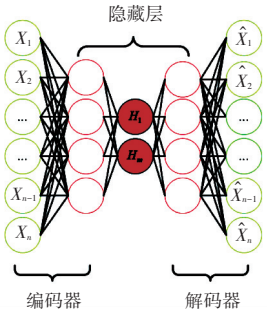

图 2　AutoEncoder 网络结构

Fig. 2　Structure of AutoEncoder network

而对于不同时间段上的微博数据，其数据量的分布不均匀可能会导致自编码器网络出现过拟合的问题，因此本文在自编码器的基础上，引入 L_2 范数进行正则化从而防止模型的过拟合问题，该过程可表示为：

$$J(W,b) = \frac{1}{n}\sum_{i=1}^{n} \| X_i - \hat{X}_i \|^2 + \frac{l}{2} \| W \|^2 \tag{9}$$

其中，n 表示输入样本个数；X_i 表示输入的高维深度文本特征向量；$\hat{X}_i$ 表示解码器重构的特征输入；W 表示自编码器网络中的参数。

1.4　KMeans++聚类分析

对降维后的深度文本特征，需要选取合适的聚类算法进行聚类分析。KMeans 聚类分析因其简单易用的特性在各类聚类分析任务中有着广泛的应用。本文选择 KMeans 算法作为深度文本特征聚类分析的算法，并通过 KMeans++对聚类过程进行优化。

对降维后的文本特征集合 $\hat{X}$，KMeans 算法通过计算最小化平方和误差（SSE）的方式来判定最佳聚类数 k 的选取，而 KMeans++对聚类初始中心点的选取过程进行了优化。在本次研究的应用场景中，KMeans++聚类算法的步骤具体如下。

步骤 1　对 $\hat{X} = \{\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_n\}$，随机选取一个点 $\hat{X}_i$ 作为初始中心点。

步骤 2　计算 $\hat{X}$ 中其他的点到 $\hat{X}_i$ 的距离 $D(x)$，并依据 $P = \dfrac{D(X)^2}{\sum_{x \in X} D(X)^2}$ 计算每个点的概率值，选取 P 最大的点作为新的聚类中心点。

步骤 3　重复步骤 2，直至获取 k 个初始中心点 $k_1, k_2, \cdots, k_n$。

步骤 4　对 $\hat{X}_m \in D$，计算 X_m 到 $k_1, k_2, \cdots, k_n$ 的距离，选取距离最近的初始中心点归类。

步骤 5　通过计算每类的均值，对 $k_1, k_2, \cdots, k_n$ 进行更新。

步骤 6　重复步骤 4、步骤 5，直至达到聚类收敛条件。

1.5　C-TF-IDF 关键词提取算法

KMeans 聚类分析过程后，得到 k 个按照深度文本特征聚类而成的文本簇。为了获取可识别理解的话题文本表示，需要对不同的话题簇进行关键词提取工作。TF-IDF 关键词提取算法通过数值统计的方式计算某个词对语料中某篇文档的重要性。将 TF-IDF 计算过程推广到聚类文本簇的关键词提取过程中，将文本簇 c 看作文档，文本簇集合看作文档集合 Doc，由此便可以推导出 C-TF-IDF 关键词权重计算方法。该方法可以表示为：

$$W_{x,c} = \| TF_{x,c} \| \times \log\left(1 + \frac{A}{f_x}\right) \tag{10}$$

其中，$TF_{x,c}$ 表示词 x 在文本簇 c 的词频 TF；f_x 表示文本簇集合 Doc 中包含词 x 的文档数；A 表示文档簇集合 Doc 的平均词数。

通过 C-TF-IDF 算法，便可以获取每个文档簇的关键词表示，并最终生成可识别理解的话题文本表示。

2　实验与分析

2.1　实验数据获取与预处理

为了验证本文提出的话题聚类分析方法的有效性，本文通过 Python 编写爬虫的方式对发生在 2019 年 4~5 月的“奔驰车主维权事件”话题相关微博进行了采集工作，数据集的统计特征见表 2。

表 2　数据集的统计特征

Table 2　Statistic feature of dataset

数据集特征	特征值
微博数	45 889
微博用户数	32 417
平均句子长度	129
平均句子词数	24

在进行话题聚类分析前，需要对采集到的微博文本进行预处理工作。文本的预处理主要涉及中文文本清洗、中文分词以及去停用词三个过程。本文通过正则表达式实现对微博文本中非文本部分、@

部分以及带有“#”的微博话题标记部分进行清洗，通过 HanLP 中文分词工具对微博文本进行分词的同时基于 TF-IDF 算法对文档库中的背景词进行筛选加入停用词表，并通过多个停用词表构建总停用词表对分词后的微博文本进行去停用词的工作。

2.2 实验环境与参数设置

本文所用的实验环境如下：CPU 为 AMD Ryzen 5900HX，内存为 32 G DDR4，GPU 为 RTX 3080 16 G。本文实验环境设置如下。

(1) 句向量的获取：本文采用 Text2Vec-base-Chinese 预训练模型进行句向量的获取，并输出 768 维的句向量。

(2) 自编码器网络：本文设计了 4 层编码器将 768 维句向量分步降维到 384、192、96、64 维，并通过对称结构的解码器重建向量到 768 维，将 64 维的句向量作为聚类分析的输入。

(3) KMeans 聚类分析：本文通过肘部法进行聚类数 k 的确定。肘部法是通过最小化点到聚类中心的距离来确定最佳的聚类数 k，该过程可以表示为下式：

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(\| x_i - \mu_j \|) \tag{11}$$

为了评测本文所提出的话题聚类分析方法，本文通过 3 个方面的指标来对话题聚类分析的结果进行展示与分析。首先，在实际的话题聚类分析过程中往往缺少带有标签的数据，因此本文采用 David-Bouldin Index(简称 *DBI* 指数) 对聚类的效果进行量化计算。*DBI* 指数的计算方法可以定义为：

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{12}$$

$$DBI = \frac{1}{k} \sum_{i,j=1}^{k} \max_{i \neq j} R_{ij} \tag{13}$$

其中，s_i 与 s_j 分别表示聚类簇 i 与 j 中所有样本点到聚类簇中心距离的平均值，d_{ij} 表示 2 个聚类簇中心的距离。通过计算 R_{ij} 可以衡量聚类簇之间的相似度。

在此基础上，式(13)通过衡量聚类簇总数为 k 的样本数据，通过最大化 R_{ij} 对应了聚类簇内的点到簇中心的距离变大，而簇之间的距离变小。因而对 *DBI* 指数来说，越大的值代表聚类簇之间的距离相近而相似，从而表示聚类效果越差，而越小的值则相反，表示更好的聚类效果。

然后，为了直观地展示各实验组的聚类结果，本文采用 UMAP 降维的方式将得到的文本特征降维至二维空间进行可视化处理。聚类可视化图通过给出不同类的特征在二维空间的位置表示，可以辅助判定聚类的紧凑度与界限，从而与 *DBI* 指标相互印证。

最后，对话题聚类分析的结果采用 C-TF-IDF 的方式进行呈现，以主观的方式评价话题聚类分析呈现的效果来评价本文话题聚类分析方法的有效性与实用性。

本文设置了 3 组 6 个对比实验在“奔驰车主维权事件”数据集上进行评测，对比实验的设置如下：

(1) 基于 Word2Vec 文本特征提取-PCA 降维的话题聚类分析方法。Word2Vec 词向量通过 gensim 工具包进行训练与获取得到 100 维的词向量，通过 PCA 降维至 64 维进行 KMeans 聚类分析。

(2) 基于 Word2Vec 文本特征提取-AutoEncoder 降维的话题聚类分析方法。通过双层 AutoEncoder 降维至 64 维进行 KMeans 聚类分析。

(3) 基于 Bert 词嵌入-PCA 降维的话题聚类分析方法。Bert 词嵌入选用哈工大讯飞联合实验室提供的 MacBert 预训练模型进行获取，输出维度为 768 维，通过 PCA 降维至 64 维进行 KMeans 聚类分析。

(4) 基于 Bert 词嵌入-AutoEncoder 降维的话题聚类分析方法。通过 4 层 AutoEncoder 降维至 64 维进行 KMeans 聚类分析。

(5) 基于 Text2Vec-PCA 降维的话题聚类分析方法。通过替换本文方法中的自编码器降维方法为 PCA 作为对照。

2.3 实验结果分析与展示

首先，从量化分析的角度对本文提出的话题聚类分析方法以及对比实验组进行量化分析。本文设定 3 组 k 值，分别为 k = 10，k = 15 以及 k = 25 进行 *DBI* 指数的量化分析以评价聚类的效果。实验结果见表 3。

表 3 不同实验组的 *DBI* 指数

Table 3 *DBI* in different group

模型	*K*		
	10	15	25
Word2Vec_PCA	2.502	2.620	2.123
Word2Vec_AE	2.069	1.997	1.957
Bert_PCA	2.139	2.260	2.244
Bert_AE	0.969	1.034	1.061
Text2Vec_PCA	3.149	3.138	3.072
Text2Vec_AE	**0.857**	**0.815**	**0.824**

从表 3 中得到的实验结果来看，静态词向量

Word2Vec 在 *DBI* 指数上的指标表现较差，这是因为 Word2Vec 模型基于给定的有限数据集语料进行训练，静态的模型结构决定了其在表示微博短文本的过程中难以应对微博中的一词多义、表意不规范以及噪声等问题，且 Word2Vec 难以捕捉动态的上下文信息，在应对复杂的文本结构时也表现不佳。

基于 Bert 词嵌入的方法在各个 k 值上的表现结果都好于 Word2Vec，这是 2 个方面的因素造成的。其一，Bert 的模型结构相较于 Word2Vec 的浅层神经网络，通过 Transformer 结构构建的深层神经网络在捕捉动态的文本特征上有着更好的表现，对多义词、文本噪声等问题也得到较好的解决；其二，相较于 Word2Vec 训练所用的小规模语料来说，Bert 预训练过程中所用到的大规模语料也使其在文本建模过程中有着更好的泛化能力，能够更好地处理不同语境下的文本特征。

本文所采用的 Text2Vec_AE 的深度文本特征提取-降维的方法在 *DBI* 指标上要好于 Bert 词嵌入的方法。Text2Vec 利用 MacBert 作为文本嵌入层，在此基础上通过孪生神经网络构建的 Bi-Encoder 结构进行文本的对比学习，从而在文本语义匹配任务上进行了优化。此外，改进的 CoSENT 损失函数通过优化预训练的目标，使得模型能够更好地区分正负样例对，也使模型的鲁棒性有所提升。

研究中注意到，实验对照组中采用自编码器结构降维的指标都显著好于 PCA 降维的对照组。这是因为 PCA 所采用的线性变换方式在处理较简单的文本特征时效果较好，而对深层神经网络所生成的非线性特征降维效果较差。从表 3 中可以看出，在应对 Word2Vec 和 Bert 等基于词嵌入的方式生成的文本向量时，PCA 尚且能够进行特征降维工作；而在基于 CoSent 的句向量表示中，PCA 降维后的特征表示效果不如 Word2Vec 以及 Bert，这也反映出 PCA 在面对复杂的非线性特征时仍有亟待改进之处。

而在采用自编码器作为降维方法的各组实验中，Bert 词嵌入与 Text2Vec 句向量表征表现较好，这是因为自编码器网络相较于 PCA 的线性变换方法，通过在不同的编解码器层之间添加非线性激活函数的方法实现了对高维文本特征中的非线性部分的有效降维。

为进一步验证本文所提出的话题聚类算法的优势，在 *DBI* 指数的基础上设定 $k = 15$，并采用 UMAP 降维的方法对得到的文本特征降维至二维平面进行可视化分析，得到的结果如图 3 所示。

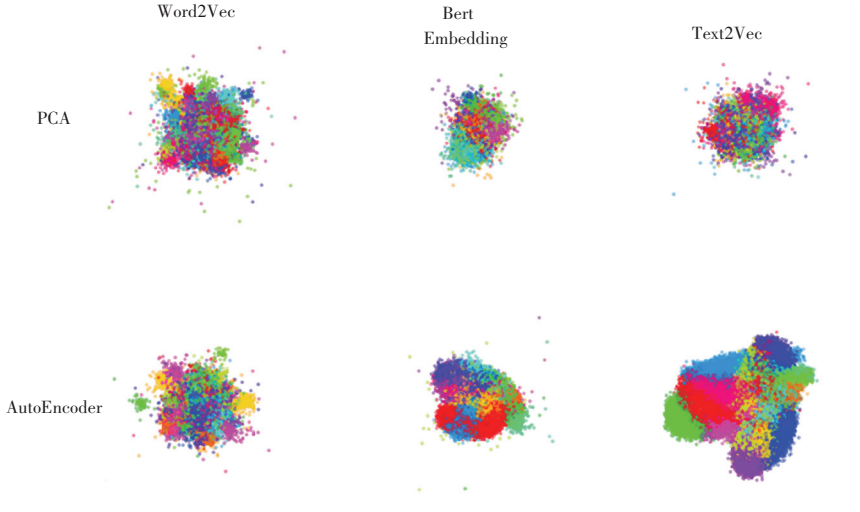

图 3　UMAP 聚类可视化图

Fig. 3　Visualization of clustering by UMAP

从聚类可视化的结果可以看出，Word2Vec 所生成的聚类簇的重叠部分较大，代表了聚类效果受到 Word2Vec 静态词向量的影响导致一词多义和表达不规范等对聚类过程产生干扰。Bert 词嵌入和 Text2Vec 的聚类可视化结果在 PCA 降维的方式中表现与 Word2Vec 相似，这是因为 Bert 和 Text2Vec 基于预训练模型的方法使得文本特征中有较多非线性部分，而 PCA 线性降维的方式难以对非线性部分进行处理，因而在 *DBI* 指数中 Bert 与 Text2Vec 相近或者劣于 Word2Vec，这也进一步验证了实验结果。最后，分析 Bert 词嵌入与 Text2Vec 通过 AutoEncoder 降维的方式生成的聚类簇可以发现，两

者能够生成更大的聚类簇,而 Text2Vec 相较于 Bert 来说生成的聚类簇更大,簇间的分界也更为明显,这也代表了 Text2Vec 在面向文本聚类的文本语义建模方面的性能要好于 Bert 词嵌入。

在此基础上,基于本文提出的话题聚类分析方法对数据集的微博话题文本进行聚类分析。首先,通过肘部法确定最佳聚类数 k, 并依据本文的方法生成聚类文本簇;随后,通过 C-TF-IDF 算法进行文本簇关键词的提取,并生成可识别的话题信息。实验结果见表 4。

在“奔驰车主维权”事件的微博话题数据集下,本文所提出的话题聚类分析方法最终生成了 17 个话题。从表 4 可以看出,本文的方法能够较好地对话题下的微博文本进行聚类分析并表示为可识别的话题关键词,具体体现为能够将对事件的评论微博(如话题 8、9、15)与对事件的报道微博(如话题 1、3、4)区分开来,在展示微博话题发展脉络的同时能够较好地反映出大众意见,从而为微博舆情的监控与引导提供支撑。

表 4 “奔驰车主维权”话题聚类分析结果

Table 4 Clustering analysis results of given topic dataset

簇编号	文本簇关键词	对应话题
1	和解,分享,回应,达成,协议	奔驰车主与奔驰达成和解
2	社会,支持,希望,道理,悲哀	微博用户表达对奔驰车主维权的支持
3	金融,服务费,收取,销售,消费	奔驰销售过程中收取金融服务费
4	暂停,运营,经销商,销售,授权	西安利之星奔驰销售被暂停授权
5	工商局,监管,品牌,中国,服务费	工商局调查金融服务费相关事项
6	兰州,引擎盖,研究生,引擎,哭诉	研究生学历的奔驰车主在引擎盖上哭诉
7	金融,销售,服务费,调查,违法	销售过程中的金融服务费涉嫌违法
8	支持,网上,逻辑,陌生人,清晰	微博用户为奔驰车主有理有据维权点赞
9	支持,合法,合理,有理有据,文化人	微博用户为奔驰车主有理有据维权点赞
10	发声,社会,学会,黑暗,希望	微博用户就奔驰店大欺客一事发表见解
11	客户,专门,歉意,沟通,回应	奔驰安排专人对接维权奔驰车主
12	车顶,看待,哭诉,公道,兰州	微博用户就奔驰车主维权事件讨论看法
13	社会,法律,真的,部门,希望	微博用户就奔驰车主维权事件延申到社会法制发表看法
14	记者,成立,联合,调查组,采访	记者报道相关部门就奔驰车主维权事件成立联合调查组
15	利益,炒作,热度,集团,受害者	部分微博用户就奔驰车主维权事件提出质疑
16	律师,名誉权,泄露,委托,纠纷	维权奔驰车主就部分用户质疑诽谤提出名誉权诉讼
17	商户,王倩,拖欠,竞集,供应商	维权奔驰车主所在公司涉嫌拖欠供应商尾款

3 结束语

本文提出的基于 Text2Vec_AE_KMeans 的话题聚类分析方法通过 MacBert 预训练模型实现了词层面的深度文本特征建模,基于 Sentence Bert 的 Bi-Enocder 结构与改进的 CosSENT 训练损失函数对文本语义表示与匹配任务进行了优化,并通过自编码器网络对高维深度文本特征进行降维,在保留非线性特征的同时减少了聚类的运算量,最后通过 KMeans 聚类算法与 C-TF-IDF 关键词提取算法实现了对文本聚类簇的关键词提取工作,并生成了可识别的话题信息。考虑到微博话题中的动态性,下一步研究拟将微博的动态过程融入到话题聚类分析的过程中,以期实现更细粒度的话题信息提取与表示,同时对聚类中的相似性较高的话题进行有效处理。

参考文献

[1] 林江豪, 周咏梅, 阳爱民, 等. 结合词向量和聚类算法的新闻评论话题演进分析 [J]. 计算机工程与科学, 2016, 38(11): 2368-2374.

[2] BÉATRICE M, CAGÉ J, HERVÉ N, et al. A French corpus for event detection on Twitter [C]// Proceedings of the 12th Language Resources and Evaluation Conference. Marseille,

France：LRA，2020：6220-6227.
[3] ZHOU Hong，LIU Jinling，WANG Xingong. Retrospective topic identification model for short text information flow[J]. Journal of Chinese Information Processing，2015，29(1)：111-117.
[4] CAO Jianping ，WANG Hui，XIA Youqing，et al. Bi-path evolution model for online topic model based on LDA[J]. Acta Automatica Sinica，2014，40(12)：2877-2886.
[5] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv，1810.04805，2019.
[6] OGUZHAN G. Deep representation learning for clustering of health Tweets[J]. arXiv preprint arXiv，1901.00439，2018.
[7] ANWAR A，ILYAS H，YAQUB U，et al. Analyzing QAnon on Twitter in context of US elections 2020：Analysis of user messages and profiles using VADER and BERT topic modeling[C]// Proceedings of the 22nd Annual International Conference on Digital Government Research. New York：ACM，2021：82-88.
[8] 刘梦颖，王勇. 基于文本双表示模型的微博热点话题发现[J]. 计算机与现代化，2021(12)：110-115.
[9] 谢梦娜. 面向微博短文本流的热点话题检测与情感社区发现方法[D]. 太原：山西大学，2022.
[10] CUI Yiming，CHE Wanxiang，LIU Ting，et al. Revisiting pre-trained models for Chinese natural language processing[J]. arXiv preprint arXiv，2004.13922，2020.
[11] REIMERS N，GUREVYCH I. Sentence-BERT：Sentence embeddings using Siamese BERT-networks[J]. arXiv preprint arXiv，2202.11456，2019.
[12] 苏剑林. CoSENT(一)：比 Sentence-BERT 更有效的句向量方案[EB/OL]. (2022-01-06). https://kexue.fm/search/词向量维度/11/.
[13] BANK D，KOENIGSTEIN N，GIRYES R. Autoencoders[J]. arXiv preprint arXiv，2003.05991，2021.