Computer Science and Application 计算机科学与应用, 2023, 13(4), 853-864
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134084

Hans汉斯

深度嵌入聚类及其在投诉文本分析中的应用

刘菲林[1]，邵立伟[2]，黄德皇[2]，喇　磊[1]

[1]对外经济贸易大学信息学院，北京
[2]中山市北京理工大学研究院，广东 中山

收稿日期：2023年3月20日；录用日期：2023年4月18日；发布日期：2023年4月27日

摘　要

针对互联网存在的巨量涉及电力投诉的用户生成超短文本，本文提出一种基于深度嵌入的聚类模型，以实现互联网电力投诉文本话题识别的方法。首先，通过改进算法进行词嵌入，以提高文本特征的语义丰度并降低数据集维度；然后，在词嵌入的基础上，借助Sentence-Bert进行句子相似度计算，从而实现短文本聚类；最后，在自主爬取的互联网用户留言中涉及电力投诉的文本数据集上部署提出的方法，完成了投诉文本的话题聚类，并与多个已有的话题识别算法在同一数据集上的效果进行比较，证明了提出模型的有效性。

关键词

词嵌入，Sententce-Bert，短文本聚类，话题识别，电力投诉

Deep Embedding Clustering and Its Application in Analysis of Complaint Text

Feilin Liu[1], Liwei Shao[2], Dehuang Huang[2], Lei La[1]

[1]School of Information Technology and Management, University of International Economics and Business, Beijing
[2]Zhongshan Research Institute, Beijing Institute of Technology, Zhongshan Guangdong

Received: Mar. 20th, 2023; accepted: Apr. 18th, 2023; published: Apr. 27th, 2023

Abstract

In view of the huge amount of Internet user-generated ultra-short text involving power complaints, a clustering model based on deep embedding is proposed to realize the topic recognition method of Internet power complaints text in this paper. Firstly, word embedding is carried out by an improved algorithm to enhance the semantic richness of text features and reduce the dimension of data set. Then, sentence similarity is calculated by using Sentence-Bert to realize short text clustering based on word embedding. Finally, the proposed method is deployed on the text data set involving power complaints

文章引用: 刘菲林, 邵立伟, 黄德皇, 喇磊. 深度嵌入聚类及其在投诉文本分析中的应用[J]. 计算机科学与应用, 2023, 13(4): 853-864. DOI: 10.12677/csa.2023.134084

in the self-crawling Internet user messages to complete the topic clustering of the complaint text, and the effect of several existing topic recognition algorithms on the same data set is compared, which proves the effectiveness of the proposed model.

Keywords

Word Embedding, Sententce-Bert, Short Text Clustering, Topic Recognition, Power Complaints

1. 引言

国家对政府公共服务能力的提升日益重视，近年来的政府工作报告和中央多个重要文件中屡次提到提升政府公共服务能力，体现了国家加快构建服务型政府的决心[1] [2] [3]。作为国民经济的支柱与命脉，电力行业与每一位公民的日常生活息息相关，公民对电力行业所提供服务的满意度，将显著影响公民对政府公共服务的总体满意度[4]。随着我国电力行业的发展，服务能力得到了显著的提升[5]，然而，当前我国电力行业的服务能力在地域上、时间上和部门上都存在不均衡现象[6]，例如在空间上，某些地区的电力服务满意度在政府各项公共服务的满意度中常年位居前列，某些地区的电力服务满意度却存在明显的问题[7]。

为了进一步提升电力行业的服务能力，不少学者开展了相关的研究。基于机器学习手段，通过对在线数据的分析来研究投诉问题、提升用户满意度是目前热门的研究方法。在全国供电服务热线 95598 开通之后，电力行业沉淀了大量的用户来电投诉的语音转文本文件。对这类文件的分析、挖掘成为电力行业满意度分析的热点。文献[8]提出了一种基于关键词提取和匹配的电力投诉文本挖掘方法，通过关键字识别和抓取，在词的层面上对投诉话题进行分类。文献[9]提出了一种基于灰度理论的相关性分析方法，并通过北京电力 95598 平台的投诉数据验证了分析方法的有效性。文献[10]提出了一种基于 K 邻近算法的电力投诉文本分类方法，并通过实验验证了该方法相对传统方法在投诉分类方面的高效性。当前，我国很多省市的政务平台都开放了用户留言与投诉功能，这些平台上面也汇聚了巨量的涉及电力的投诉信息。关键是，这类平台是开放数据获取平台，任何用户都可以访问其他用户的投诉内容，因而这类平台中的负面消息的影响力更大[11]。并且根据研究，开放平台中的用户投诉具有跟风与放大效应[12] [13]。因此，对开放投诉平台上在线投诉的分析，对于提升电力行业服务能力、消除相关负面影响具有重大意义。

对于巨量在线投诉文本的分析，文献[14]提出了一种基于情感概率主题模型的在线投诉挖掘方法，该方法需要借助用户打分(评星)数据辅助提升分析结果，而政府公开投诉平台往往没有打分选项。基于聚类的在线投诉分析是一类流行的方法[15] [16] [17]，然而政府留言平台的用户留言文本往往很短，聚类方法直接应用于超短文本的话题分析往往存在高维稀疏导致的类别区分度低、无法支撑下游应用等问题[18]。还有些研究基于人工建立的领域词库，在餐饮业[19]、酒店业[20]等专用领域的投诉分析中取得了不错的效果，然而领域词库的建立耗费大量的人工开销，且难以迁移到其他领域。词嵌入技术可以很好地降低数据集的特征维度，对短样本的聚类性能具有显著的提升[21]。针对传统方法在互联网超短文本分析中所面临的维度灾难、领域词典和专家经验依赖以及类别差异度低等问题，本文在词聚类的基础上，提出了

一种改进的 Word2Vec 方法，并基于此提升聚类性能，构建完整的文本聚类与投诉话题识别框架。通过对互联网爬取的涉及电力的巨量投诉文本的分析，成功识别出投诉的热门领域和主题，为提高电力服务能力和行业形象提供有力工具。

2. 词嵌入与 Word2vec

词嵌入是一种把代表文档集中所有词本身的高维空间映射到一个维度被大大降低的连续向量空间的过程，映射的对象根据需要可以是字、单词或者短语。通过词嵌入可以起到特征降维的效果，更重要的是，降维之后的特征可能具有更好的语义丰度，有助于提升下游应用的性能。词嵌入原理如图 1 所示。

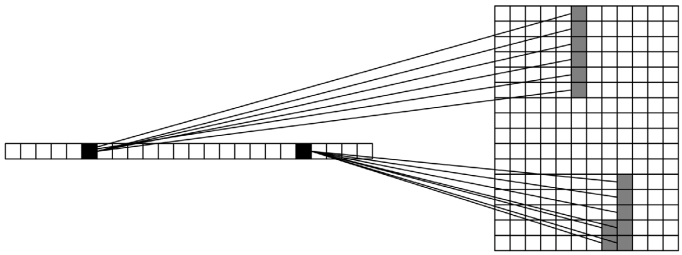

Figure 1. Principle of word embedding
图 1. 词嵌入原理

词嵌入曾经以 one-hot 编码等方式作为主流手段，随着深度学习的兴起，基于神经网络的词嵌入逐渐成为热门方法。

2.1. Word2Vec

Word2Vec 是一种主流的基于神经网络的词嵌入方法。它通过文本集的上下文来推断单词，假设数据集的所有单词为$\{w_0, w_1, \cdots, w_{N-1}\}$，则 Word2Vec 的目标是对于上下文单词的预测使式(1)概率最高。

$$P\left(w_{i-m}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+m} \middle| w_i\right) = \prod_{j\neq i,\, j=1-m}^{i+m} P\left(w_j \middle| w_i\right) \tag{1}$$

在式(1)中，为了使等式右边成立，需要词之间的独立假设。设词汇量大小为 V，词嵌入维度为 D，则可定义一个 $V*D$ 矩阵来存储词嵌入。将每个词表示为大小为 V 的 one-hot 编码向量，由于输入单词和相应的输出单词各自的大小为 V，设第 i 个输入的词为 x_i，对应的嵌入为 z_i，输出为 y_i，对于它们的对应关系，可通过式(2)和式(3)表示：

$$\text{logits}\left(x_i\right) = z_i W + b \tag{2}$$

$$\hat{y}_i = \text{softmax}\left(\text{logits}\left(x_i\right)\right) \tag{3}$$

其中，logits(x_i)表示 x_i 的非标准化分数，W 是 $D*V$ 权重矩阵，b 是偏置矢量，$\hat{y}_i$ 表示输出的预测。通过对输入单词的计数，可以使用符对数似然损失函数来计算给定样本点(x_i, y+i)的损失。损失函数如式(4)：

$$J\left(\theta\right) = -\left(\frac{1}{N-2m}\right)\sum_{i=m+1}^{N-m}\sum_{j\neq i,\, j=i-m}^{i+m} \log P\left(w_j \middle| w_i\right) \tag{4}$$

由于 w_i 采用 one-hot 编码，且 $P(w_j|w_i)$可由第 n 个数据点给出，所以存在式(5)：

$$P\left(w_j \middle| w_i\right) = \frac{\exp\left(\text{logits}\left(x_n\right)_{w_j}\right)}{\sum_{w_k \in vocabulary} \exp\left(\text{logits}\left(x_n\right)_{w_k}\right)} \tag{5}$$

其中，$\text{logits}(x_n)_{w_j}$ 表示 w_j 的 one-hot 编码中非零的索引所对应的得分值。因此，可以得到有效损失函数的简化形式如式(6)：

$$J(\theta)=-\frac{1}{N}\sum_{i=m+1}^{N-m}\sum_{j\neq i,\,j=i-m}\left(\sigma\left(\text{logits}(x_n)_{w_j}\right)\right)+\sum_{q=1}^{k}E_{w_q vocabulary-(w_i,w_j)}\log\left(\sigma\left(\text{logits}(x_n)_{w_q}\right)\right) \tag{6}$$

其中，w_j 表示 w_i 的上下文单词，w_q 表示其非上下文单词，σ 表示 sigmoid 激活函数。为了使式(6)最小化，应该使 $\sigma\left(\text{logits}(x_n)_{w_j}\right)\approx 1$。

对于单词表征分布的学习过程，负采样是理想的方式。单词表征是通过预测一个训练词周围的单词来学习的。在训练时，正确的周围单词提供积极的例子，而不是一组抽样的负样本(噪音)。为了找到这些负样本，噪声分布可定义如式(7)：

$$Pn(w)=U(w)^{\frac{3}{4}}\Big/\sum_{i=1}^{|vocabulary|}U(w_i)^{\frac{3}{4}} \tag{7}$$

2.2. 改进的 Word2Vec

通常对 Word2Vec 的负采样都采用了式(7)的形式，然而如式(7)的单字分布只考虑词频，对不同的目标词选择负采样时，只能提供相同的噪声分布，这样是无法反映文本的内容信息的。为此，本文提出了一种基于单词共现的负采样策略。

词共现网络是一种表示词在文本集中共现关系的关系图模型。设数据集内的单词为顶点，当两个单词同时出现在一个句子中时，与传统负采样只考虑邻接词不同，根据它们的距离，可以创建一个无向边，只要它们的距离不超过参数 $d_{\max}$。单词间的距离定义如式(8)：

$$d(w_a,w_b)=|j-i| \tag{8}$$

其中，i、j 代表单词 w_a 和 w_b 在句子中出现的位置(句子中以词为单位的序号)。对于 w_a 和 w_b，当距离为 λ 时，它们的共现度为：

$$cooc(\lambda,w_a,w_b)=\left|\left\{(w_a,w_b)\middle|d(w_a,w_b)=\lambda\right\}\right| \tag{9}$$

对于边(w_a, w_b)的权重，可以通过 w_a 和 w_b 之间满足 $\lambda<d_{\max}$ 的总共现度来定义，如式(10)：

$$weight(d_{\max},w_a,w_b)=\sum_{\lambda=1}^{d_{\max}}cooc(\lambda,w_a,w_b) \tag{10}$$

无向加权词共现网络也可以用|W|*|W|阶的对称邻方阵 $\boldsymbol{A}$ 表示，其中 W 是文档集的词总数。对邻接矩阵 $\boldsymbol{A}$ 按行进行归一化处理，将其转化为右随机矩阵 $\boldsymbol{S}$。对于负采样，可以统一地随机为每个训练单词抽取一个它周围的单词作为(正向的)目标单词，这个范围由训练文本 c 的大小决定。也就是说，训练词 w_t 的周围词 w_s 必须满足 $d(w_t,w_s)\le c$。对于相同的文本集，设 $c=d_{\max}$，邻接矩阵中的元素 S_{ab} 代表词 w_a 被选作训练词 w_b 的目标单词的概率。因此，行 S_a 显示训练词 w_a 在训练整个数据集后的目标词分布情况，并且无论在语料库上进行了多少次训练迭代，这个分布都不会改变。因此，可得式(11)：

$$P_{bigram}(w_a,w_b)=\frac{\sum_{\lambda=1}^{d_{\max}}cooc(\lambda,w_a,w_b)}{\sum_{i=1}^{|vocabulary|}\sum_{\lambda=1}^{d_{\max}}cooc(\lambda,w_a,w_i)}=S_{ab} \tag{11}$$

根据以上步骤，可通过对词共现的统计生成一个词关系网络结构。在生成词共现网络之后，可以应用随机游走得到另一个负采样噪声分布矩阵。对于词共现网络上的随机游走，设开始于一个初始顶点 w_a，在每一步中可以通过 w_a 的一条边到达另一个顶点，记作 w_b。对于一个带权的词共现网络，可以定义从顶点 w_a 到 w_b 的转移概率为 $P(w_a, w_b)$，它实际代表了边(w_a, w_b)的权重在与 w_a 相连的所有边的总权重中的占比。通过邻接矩阵 $\boldsymbol{A}$ 和随机矩阵 $\boldsymbol{S}$，转移概率可以根据式(12)表示：

$$P\left(w_a, w_b\right) = A_{ab} \Big/ \sum_{i=1}^{|A_u|} A_{ai} = S_{ab} \tag{12}$$

对于学习所有训练词的转移概率，可以在所有顶点上应用随机游走，使每个训练词同时成为一个 t 步随机游走的初始顶点。整个转移概率集可以表示为一个转移矩阵，在本文中，这正是词共现生成网络的右随机矩阵 $\boldsymbol{S}$。我们发现矩阵 $\boldsymbol{S}$ 中的自循环(开始和结束于同一顶点的边：邻接矩阵或随机矩阵的主对角线)代表了单词在其自身上下文中的出现，这种情况可能会重复出现。我们假设它们构成伪事件，因此在矩阵 $\boldsymbol{S}$ 中自循环被移除。为了观察自循环的效果，我们对矩阵 $\boldsymbol{S}$ 进行 t 步随机游走。在此基础上，t 步随机游走转移矩阵的元素可由式(13)表示：

$$P_{\text{randomwalk}}\left(w_a, w_b\right) = S_{ab}^{t} \tag{13}$$

词共现网络是高度连通的，对于这样的网络，随机游走只需几步就可以收敛到稳定状态。稳态意味着无论从哪个顶点开始，目标顶点概率的分布都保持不变。换句话说，所有 S_t 列都将具有相同的值。所以，我们将最大步数 $t_{\max}$ 设置为 4。我们将使用这些 t 步随机游走转移矩阵作为负采样噪声分布矩阵的基础。

从基本噪声分布矩阵出发，可以利用幂函数对分布进行调整。通过规范化这个调整后矩阵的所有行，使每一行之和为 1，可得式(14)：

$$P_n\left(w_a, w_b\right) = \left(B_{ab}\right)^{P} \Big/ \sum_{i=1}^{|B_u|} \left(B_{ai}\right)^{p} \tag{14}$$

其中，B 是通过式(11)和(13)求得的基本噪声分布，p 代表幂次。通过以上步骤，在负采样 Word2Vec 训练时，对于每个训练词，我们使用噪声分布矩阵中相应的行来代替原来的一元组噪声分布来选择噪声样本候选。

3. 基于词嵌入的聚类与话题识别

3.1. Sentence-Bert

传统的聚类方法用于在线短文本时，常常存在聚出的类簇区分度不高，或者缺乏语言学意义的问题。通过 Bert 等词向量模型进行特征聚类，然后再用聚类算法处理聚簇之后的特征，或者直接用词向量模型聚类，是目前新兴的方法。

在嵌入阶段，Bert 通过双向编解码器块连接，目标函数为 $P\left(w_{\xi} \middle| w_1, \cdots, w_{\xi-1}, w_{\xi+1}, \cdots, w_n\right)$，Bert 模型结构如图 2 所示。

虽然基于 Bert 的聚类，类簇间的区分度有所提升，但时间开销巨大[22]。Sentence-Bert 避免了经典 Bert 算法需要将句子两两输入网络计算的问题，从而大大提升了时间性能。

Sentence-Bert 在 Bert 的输出中添加了一个池化操作，以导出固定大小的句子嵌入。在 Sentence-Bert 中，默认的池化策略是计算所有输出向量的平均值，为了微调 Bert，可以创建三元组网络来更新权重，以便生成的句子嵌入在语义上有意义，并且可以与余弦相似性进行比较。网络结构取决于可用的训练数据。

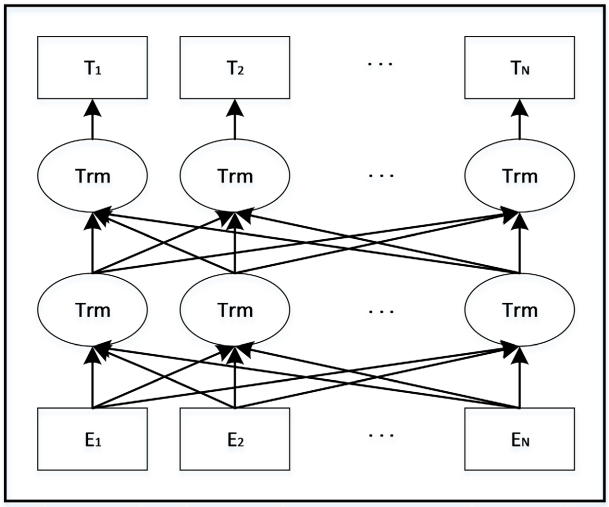

Figure 2. Structure diagram of Bert
图 2. Bert 结构图

对于 Sentence-Bert 的分类目标函数，我们将嵌入了 a 和 b 的句子与元素级差异 $|a-b|$ 连接起来，然后再乘以可训练的权重 W_t，如式(15)：

$$o = \text{softmax}\left(W_t\left(a, b, |a-b|\right)\right) \tag{15}$$

其中，$W_t \in \mathbb{R}^{3n\times k}$，$n$ 是句子嵌入的维度，k 是标签数量，我们优化了交叉熵损失，如图 3 所示。

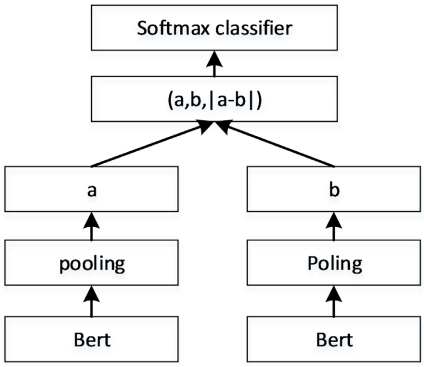

Figure 3. Classification objective function structure of Sentence-Bert
图 3. Sentence-Bert 的分类目标函数结构

对于 Sentence-Bert 的回归目标函数，可以通过计算嵌入了 a 和 b 的两个句子的余弦相似度来求解。在这里使用均方误差损失作为目标函数，回归目标函数结构如图 4 所示。

对于 Sentence-Bert 的三重态目标函数，给定一个锚定句 a，一个正句 p，一个负句 n，三元组损失可以调整网络，使得 a 和 p 之间的距离小于 a 和 n 之间的距离。在数学上，需要最小化以下损失函数：

$$\max\left(\left\|s_a - s_p\right\| - \left\|s_a - s_n\right\| + \varepsilon, 0\right) \tag{16}$$

其中，s_x 表示 $a/n/p$ 的句子嵌入，$\|\cdot\|$ 表示距离，在这里使用欧氏距离，边距 ε 确保 s_p 比 s_n 更接近 s_a。

通过以上步骤，可以通过句子相似度和句子级别的嵌入来实现短文本聚类。

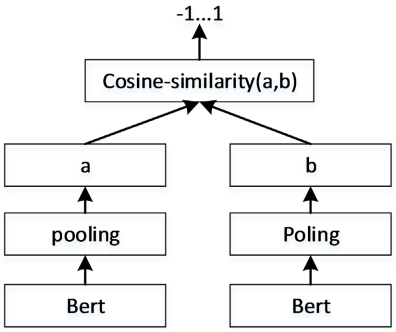

Figure 4. Regression objective function structure of Sentence-Bert
图 4. Sentence-Bert 的回归目标函数结构

3.2. 话题识别全流程

为了验证本文提出方法的有效性，我们开发了一个电力投诉文本获取与分析系统，并将本文提出的算法部署于该系统中。在数据获取方面，针对目标网站，我们应用 selenium + ChromeDriver 方法和 request 解析方法开发了定制化的爬虫，目标网站为各省市电力网站留言板，或各省市政务网站留言板块中涉及电力投诉的文本。以河北新闻网阳光理政栏目的电力留言板块为例，目标数据如图 5 所示。

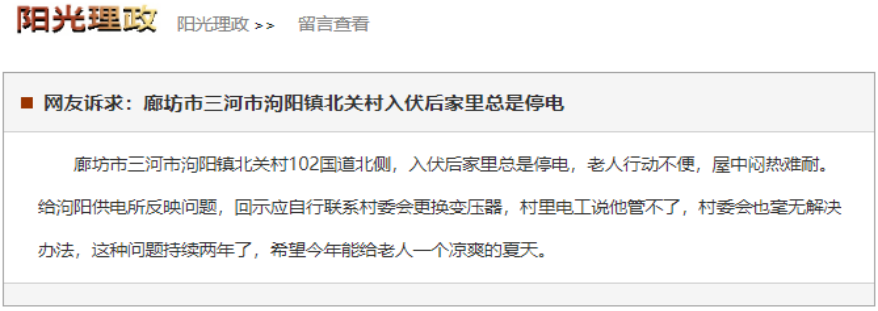

Figure 5. Example of objective source data in this study
图 5. 本文目标源数据示例

传统的中文自然语言处理任务，包括数据清洗与降噪、去除停用词以及分词。然而，对于 Bert 等深度嵌入模型，以字作为处理对象往往可以达到比分词之后再分析更优的结果[23]，并且可以节约时间开销。因此，本文的预处理主要是降噪与去除停用词。在线评论类文本除了可以用正则表达式过滤的噪声，往往还包括太短的、缺少语义信息的文本以及不相关的文本。为此，我们设置了长度阈值以删除过短文本，并通过关键词过滤删除不相关的文本。在去除停用词方面，我们借助哈工大停用词表和百度停用词表。基于深度嵌入聚类的话题识别算法流程如图 6 所示。

通过如图 5 所示的步骤，可以在词嵌入的基础上降低文本集特征维数并提升语义丰度，从而通过 Sentence-Bert 可以实现更好的聚类。对于聚类的结果打上话题标签，从而可以实现涉及电力投诉文本的话题输出。

4. 实验与结果分析

在实验部分，我们通过自主开发的爬虫获取了河北阳光理政网、长江网武汉城市留言板、湘问投诉直通车等省市的政务网站或公共留言网站，以及各市 12345 在线留言网站中 2016 年 1 月 1 日至 2020 年 6 月 30 日涉及电力投诉的文本 115326 条，去除无效信息后的目标文本 90035 条。

算法：基于深度嵌入聚类的电力投诉文本的话题识别算法

输入：爬取的投诉文本集S、word2vec及Sentence-Bert参数、相似度阈值ρ，长度阈值L_{min}

输出：涉及电力投诉的各类话题标签及对应文本

```
正则表达式降噪
if (文本长度<)L_min删除
else 加入预处理文本集
去除停用词
特征词嵌入
for (i=0; i<N; i++)
    if (相似度<ρ)
        形成新簇
    Else
        加入应有簇
for (j=0; j<C; j++)
    类簇j打类别标签
```

Figure 6. Topic recognition based on deep embedding clustering
图 6. 基于深度嵌入聚类的话题识别

4.1. 外部指标评价

作为一种无监督学习，聚类算法的效果评价相对复杂，难以使用分类中准确率、F_1 之类的直观指标来衡量。聚类效果的评价指标一般包括内部评价、外部评价和相关评价三大类，其中外部评价是相对更客观的评价指标，目前更常用于聚类效果评价。

标准化互信息(Normalized Mutual Information, NMI)是一种经典的基于信息论评价两类分布吻合程度的聚类外部评价指标，定义如式(17)：

$$\mathrm{NMI}(\Omega,C)=\frac{I(\Omega;C)}{\left(H(\Omega)+H(C)/2\right)} \tag{17}$$

其中，I 表示互信息，H 表示熵，该指数值介于(0, 1)，值越大效果越好。本文以爬取并预处理后的 9 万多个文本为测试数据集，对比了本文提出方法和其他对照算法的 NMI 指标，实验结果如表 1 所示。

Table 1. Comparison of clustering NMI index among algorithms
表 1. 各算法聚类 NMI 指标对比

算法	平均	最大	最小
sk-means	0.307	0.342	0.289
SDEC	0.516	0.544	0.501
KB-DBSCAN	0.479	0.505	0.436
Birch	0.505	0.552	0.464
深度嵌入聚类	0.697	0.713	0.680

如表 1 所示，实验部分采用的对照算法包括 sk-means [15]、SDEC [16]以及 KB-DBSCAN [17]，这些算法都是在 K-means、DBSCAN、神经网络等经典聚类算法基础上的改进算法。与对照算法相比，本文提出的深度嵌入聚类具有最优的 NMI 值。

兰德指数也是聚类中经典的外部评价指标，它将聚类理解为一系列决策过程，并评价正确决策的比例。目前在聚类结果评价中，常用调整的兰德指数(Adjusted Rand Index, ARI)，如式(18)所示：

$$\mathrm{ARI}=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\Big/\binom{n}{2}\right]}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\Big/\binom{n}{2}\right]} \tag{18}$$

其中，n_{ij} 表示同时位于簇 x 与簇 y 中的样本数量，a_i 表示 x_i 中的样本数量，b_j 表示 y_j 中的样本数量，n 表示总的样本数量。实验结果如表 2 所示。

Table 2. Comparison of clustering AMI index among algorithms
表 2. 各算法聚类 AMI 指标对比

算法	平均	最大	最小
sk-means	0.511	0.539	0.495
SDEC	0.644	0.668	0.629
KB-DBSCAN	0.919	0.928	0.905
Birch	0.877	0.896	0.854
深度嵌入聚类	0.912	0.920	0.903

如表 2 所示，本文提出的算法具有次优的 ARI 指标，并且与最优的算法差距很小，综合 NMI 和 ARI 指标，可以认为本文提出算法的聚类效果是最优的。

4.2. 时间开销分析

时间开销是在线 UGC 分析重要的指标，因为在线 UGC 往往是巨量的，过高的时间开销难于支撑实时在线应用。我们对比了本文提出算法和对照算法处理相同数据量时的时间开销，结果如图 7 所示。

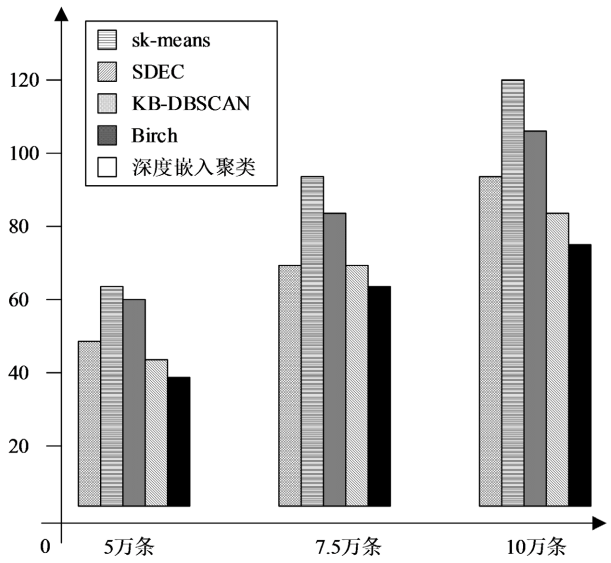

Figure 7. Time consumption of each algorithm
图 7. 各算法时间开销

如图 7 所示，本文提出的算法具有较理想的时间开销特性，对于巨量互联网数据的实时分析，低时间开销对支撑业务上线具有重要意义。

4.3. 话题识别结果

通过对电力投诉文本的聚类结果打标签，我们获得了投诉文本的话题识别结果。在打标签的过程中，我们剔除了文本数量偏少的小部分类簇，剩余类簇的话题标签结果如表 3 所示。

Table 3. Topic recognition result of power complaints
表 3. 电力投诉话题识别结果

话题	例文
费用纠纷	普通居民用电电表一个月走了 3500 多度，电工按表收费，不掏钱就断电……
停电断电	小区不知道什么原因停电，村里已经好些天没电。但不知原因……
农村用电	沙河供电公司拉闸限制农业用电，导致无法灌溉……
基础设施	电压低，售后检查电压 170 即便是上午空调也是制冷效果差，电脑总死机……
服务态度	多次反映问题接待人员推诿敷衍，后来态度冷漠，不理我们……
安全隐患	一农户家的菜秧子竟然蹿到了房顶的线路上，而且盘根错节，让人害怕……
物业阻碍	九方物业不卖给业主生活用电，经业主委员会多次沟通都没有结果……

在表 3 中，话题按照类簇中的文本数量排序，可见涉及电力投诉最多的话题是费用纠纷和停电断电，此外，涉及农村用电的投诉话题也较多。以涉及安全隐患的投诉话题为例，该话题涉及的关键词词云如图 8 所示。

Figure 8. Word cloud of potential safety hazard
图 8. 安全隐患话题词云

虽然安全隐患类投诉在涉电投诉话题中占比并不高，但安全无小事，通过话题识别，话题内的文本和词云可以了解到，配电箱、电线杆和变压器是隐患高发设备，隐患类型主要是漏电、设备倾斜以及违章堆放造成的火灾风险。对于其他话题，也可以根据本文提出的改进的 Word2Vec 进行词聚类分析，从而识别出每类话题中的关键因素、高风险因素。

5. 结论

通过机器学习方法分析用户在线留言、在线投诉以提高营销与服务水平是大数据与移动互联网时代各行业流行的方法。在电力行业领域，目前相关研究开展得并不多。本文提出了一种基于词嵌入的在线短文本聚类与话题识别模型。该模型首先通过改进的 Word2Vec 算法进行特征聚类，以降低数据集维度

并提高特征的语义丰度，然后在特征嵌入的基础上，通过 Sentence-Bert 实现话题聚类，最后通过对类簇打标签的方式实现话题识别。实验结果表明，本文提出的模型可以有效地对在线涉电投诉进行话题聚类与识别，相对其他对照算法，本文提出的算法具有更好的聚类性能指标和更低的时间开销。话题识别的结果可用于提高电力行业的服务能力和口碑，根据投诉主题进行有针对性的服务提升还可以有效避免互联网投诉的跟风与聚集效应，提升电力企业的企业形象。

电力行业通过 95598 沉淀了大量的投诉语音与文本信息，并且通过投诉热线所描述的问题一般是长文本，如何将 95598 后台的长文本和互联网平台的前台短文本融合起来，进行一体化的投诉话题挖掘，是这一领域未来值得投入精力的研究方向。

基金项目

中山市科技计划项目基于人工智能 CT 时序列的肺癌早期预测及其应用(2019AG009)。

参考文献

[1] 张檣, 程倩. 服务型政府的知识建构与扩散——基于 SKAD 的 5T 话语分析[J]. 学习论坛, 2020(4): 46-52.

[2] 李晓飞. 户籍分割、资源错配与地方包容型政府的置换式治理[J]. 公共管理学报, 2019, 16(1): 16-28.

[3] 丁志刚, 王杰. 中国行政体制改革四十年: 历程、成就、经验与思考[J]. 上海行政学院学报, 2019, 20(1): 35-47.

[4] Gencer, B., Larsen, E.R. and van Ackere, A. (2020) Understanding the Coevolution of Electricity Markets and Regulation. *Energy Policy*, **143**, Article ID: 111585. https://doi.org/10.1016/j.enpol.2020.111585

[5] 胡洋, 田兵, 雷金勇, 等. 面向能源互联的分布式发电系统聚合服务运营模式分析[J]. 中国电力, 2020, 53(8): 1-8.

[6] 朱州. 基于大数据分析的电力客户服务需求预测[J]. 沈阳工业大学学报, 2020, 42(4): 368-372.

[7] 冷媛, 陈政, 黄国日, 等. 偏远山区电力普遍服务微网优化模型研究[J]. 智慧电力, 2020, 48(6): 61-66.

[8] 刘志欣, 黄旭, 魏加项, 等. 基于 95598 大数据的电力客户满意度分析[J]. 电力大数据, 2018, 21(8): 19-24.

[9] Liu, Z.X., Huang, Z., Yu, L., Meng, C. and Zhou, J.Q. (2018) Power Customer Complaints Model Based on Grey Correlation Analysis Method. 2018 2*nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference* (*IMCEC*), Xi'an, 25-27 May 2018, 1411-1415. https://doi.org/10.1109/IMCEC.2018.8469242

[10] 吴艾薇, 雷景生. 面向电力客户投诉信息的短文本分类算法的改进技术[J]. 上海电力学院学报, 2017, 33(6): 597-600.

[11] Huang, D., Wang, C.-D., Wu, J.-S., Lai, J.-H. and Kwoh, C.-K. (2019) Ultra-Scalable Spectral Clustering and Ensemble Clustering. *IEEE Transactions on Knowledge and Data Engineering*, **32**, 1212-1226. https://doi.org/10.1109/TKDE.2019.2903410

[12] Yang, X., Li, G.X. and Huang, S.S. (2017) Perceived Online Community Support, Member Relations, and Commitment: Differences between Posters and Lurkers. *Information & Management*, **54**, 154-165. https://doi.org/10.1016/j.im.2016.05.003

[13] 杨东红, 吴邦安, 陈天鹏, 薛红燕. 基于京东商城评价数据的在线商品好评、中评、差评比较研究[J]. 情报科学, 2019, 37(2): 125-132.

[14] Li, X.L., Wu, C.J. and Mai, F. (2019) The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis. *Information & Management*, **56**, 172-184. https://doi.org/10.1016/j.im.2018.04.007

[15] Moshtaghi, M., Bezdek, J.C., Erfani, S.M., Leckie, C. and Bailey, J. (2019) Online Cluster Validity Indices for Performance Monitoring of Streaming Data Clustering. *International Journal of Intelligent Systems*, **34**, 541-563. https://doi.org/10.1002/int.22064

[16] Ren, Y.Z., Hua, K.R., Dai, X.Y., *et al*. (2019) Semi-Supervised Deep Embedded Clustering. *Neurocomputing*, **325**, 121-130. https://doi.org/10.1016/j.neucom.2018.10.016

[17] Chen, Y.W, Zhou, L.D, Pei, S.W, *et al*. (2019) KNN-BLOCK DBSCAN: Fast Clustering for Large-Scale Data. *IEEE Transactions on Systems, Man, and Cybernetics*: *Systems*, **51**, 3939-3953. https://doi.org/10.1109/TSMC.2019.2956527

[18] Chen, M.S., Huang, L., Wang, C.D. and Huang, D. (2020) Multi-View Clustering in Latent Embedding Space. *The Thir-*

ty-Fourth AAAI Conference on Artificial Intelligence, New York, 7-12 February 2020, 3513-3520.

[19] Giglio, S., Bertacchini, F., Bilotta, E. and Pantano, P. (2019) Using Social Media to Identify Tourism Attractiveness in Six Italian Cities. *Tourism Management*, **72**, 306-312. https://doi.org/10.1016/j.tourman.2018.12.007

[20] 孙长伟, 任宗来, 杨俊杰, 庞坤亮. 基于评论数据的酒店服务质量的细粒度分析[J]. 计算机应用与软件, 2019, 36(7): 32-38.

[21] Kim, S., Park, H. and Lee, J. (2020) Word2vec-Based Latent Semantic Analysis (W2V-LSA) for Topic Modeling: A Study on Blockchain Technology Trend Analysis. *Expert Systems with Applications*, **152**, Article ID: 113401. https://doi.org/10.1016/j.eswa.2020.113401

[22] Qu, C., Yang, L., Qiu, M.H., *et al*. (2019) BERT with History Answer Embedding for Conversational Question Answering. *Proceedings of the* 42*nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, 21-25 July 2019, 1133-1136. https://doi.org/10.1145/3331184.3331341

[23] Gao, Z.J., Feng, A., Song, X.Y. and Wu, X. (2019) Target-Dependent Sentiment Classification with BERT. *IEEE Access*, **7**, 154290-154299. https://doi.org/10.1109/ACCESS.2019.2946594