Computer Science and Application 计算机科学与应用, 2024, 14(2), 480-488
Published Online February 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.142048

Hans汉斯

基于文本主题和地理位置的生活日志分类方法

贾智颖

沈阳建筑大学计算机科学与工程学院，辽宁 沈阳

收稿日期：2024年1月26日；录用日期：2024年2月22日；发布日期：2024年2月29日

摘　要

我们从2011年开始，通过开发的App有计划地收集个人生活日志数据，目前已经有22位志愿者参与到这个项目中，收集到的有效生活日志数据超过4万余条。将这些丰富而杂乱的数据进行分类，为人们提供更清晰、有序的生活见解是一件有意义的事情。本文提出了一个生活日志文本分类模型DTC-TextCNN，通过引入LDA主题模型，对文本日志的主题特征进行提取；使用DB-SCAN算法，对发送动态时的地理位置进行聚类，得到不同的地理位置特征簇，并将提取到的文本主题特征和地理位置特征与文本动态进行拼接，输入到TextCNN模型中进行分类。实验结果表明，将地理位置这一特征引入模型中，有助于更好地理解文本发生的背景和环境，提供更丰富的上下文信息。融合了地理特征和主题特征的分类方法，弥补了生活日志文本语义模糊以及全局语义缺失的问题，提高了对于文本内容的理解水平。通过在Liu Lifelog数据集上的测试，可以看到该模型能够提高对生活日志分类的准确性。

关键词

生活日志，深度学习，文本分类

Lifelog Classification Method Based on Text Theme and Geographic Location

Zhiying Jia

School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang Liaoning

Received: Jan. 26th, 2024; accepted: Feb. 22nd, 2024; published: Feb. 29th, 2024

Abstract

We have been systematically collecting personal lifelog data through the development of an app

文章引用: 贾智颖. 基于文本主题和地理位置的生活日志分类方法[J]. 计算机科学与应用, 2024, 14(2): 480-488.
DOI: 10.12677/csa.2024.142048

since 2011. Currently, 22 volunteers have participated in this project and have collected over 40000 effective lifelog data. Classifying these rich and chaotic data to provide people with clearer and more organized insights into their lives is a meaningful thing. This article proposes a lifelog text classification model, DTC-TextCNN, which extracts topic features from text logs by introducing the LDA topic model; Using the DB-SCAN algorithm to cluster the geographical locations when sending dynamics, obtain different geographical feature clusters, concatenate the extracted text topic and geographical location features with the text dynamics, and input them into the TextCNN model for classification. The experimental results indicate that incorporating the feature of geographic location into the model helps to better understand the background and environment of text occurrence, providing richer contextual information. The classification method that integrates geographical and thematic features compensates for the problems of semantic ambiguity and global semantic loss in life log texts, and improves the level of understanding of text content. Through testing on the Liu Lifelog dataset, it can be seen that the model can improve the accuracy of lifelog classification.

Keywords

Lifelog, Deep Learning, Text Classification

1. 引言

近年来，随着社交软件的普及，人们开始在微博、Twitter 和 Instagram 等社交媒体上用日记记录事件[1]，同时由于传感技术和位置感知技术的进步，使得人们对事件的记录更加方便准确，记录的信息也更加丰富[2]。因而对于生活日志的分析已经成为一个重要的研究方向。目前，Lifelog 已经被应用于许多领域的研究中。在医疗方面，Lifelog 已被用于查看肥胖患者的行为变化，以进行体重管理[3]；使用抑郁症患者的 Lifelog 数据来预测他们抑郁症复发的风险[1]；通过 Lifelog 查看患者健康状态[4]。在社会生活方面，利用 Lifelog 中记录的位置信息对个人的移动情况进行分析[5] [6]；使用地理标记照片数据集识别用户的重要位置和日常行为[7]；使用个人信息型生活日志数据自动生成故事模型[8]。Lifelog 数据内容丰富多样，为了提高管理效率，对其进行合理分类显得尤为必要。Lifelog 中包含大量对日常行为的描述，根据这些行为对其分类是一个很好的分类方式。通过分类，我们能更有条理地整理个体的日常活动。

目前，常用的文本分类机器学习算法主要有朴素贝叶斯(NB) [9]，K 最近邻(KNN) [10]和支持向量机(SVM) [11]。但这些方法忽略了文本数据中的上下文信息，导致语义信息无法准确表达，同时存在时间成本较高等问题，影响分类效率。近年来基于深度学习的文本分类方法引起了人们的广泛关注，Kim 等人提出了 TextCNN 的分类模型[12]，该模型采用卷积操作对文本局部特征进行提取，取得了不错的效果；王佳慧等[13]将 CNN 与 Bi-LSTM 混合模型，有效提升了中文文本分类准确性；杨阳等采用融合词向量的方法来提高文本分类精度[14]；AK Sharma 等人通过融合 Word2Vec 技术，同时对 CNN 模型进行微调来提高分类的准确性[15]。由于生活日志是人们随手对生活的记录，因此存在词汇不规范、以及语义模糊的问题，现有的文本分类模型大多基于文本内容本身，通过提高挖掘文本语义能力来提高文本的分类效果，但这对于语义表达不规范的生活日志来说分类效果的提升是有限的。因此本文提出了融合地理位置特征和主题特征的生活日志文本分类模型 DTC-TextCNN，通过利用用户发送动态的地理位置这一空间信

息，从而更为准确的表示出句子的语义，在一定程度上弥补了语言表达模糊以及不规范的问题，同时引入主题模型，弥补了 CNN 模型对于全局文本语义缺失的问题。

文章的组织结构如下：第一节介绍了 Lifelog 的发展以及常用的分类方法，第二节介绍了我们的 Lifelog 项目及使用的数据集，第三节介绍了 DTC-TextCNN 模型的构建方式，第四节对模型结果进行评价，最后进行了总结和评论。

2. Lifelog 项目和数据集

2.1. Lifelog 项目

我们团队从 2011 年开始有计划地收集个人 Lifelog 数据。目前已经有 22 位志愿者参与到这个项目中，在过去的 11 年，几乎每一天都会记录一次数据，收集到的有效 Lifelog 数据超过 4 万条。志愿者可以登录我们团队开发的 App，随时随地记录个人生活，存留下每一个精彩瞬间。图 1 是我们的开发的 APP 界面。

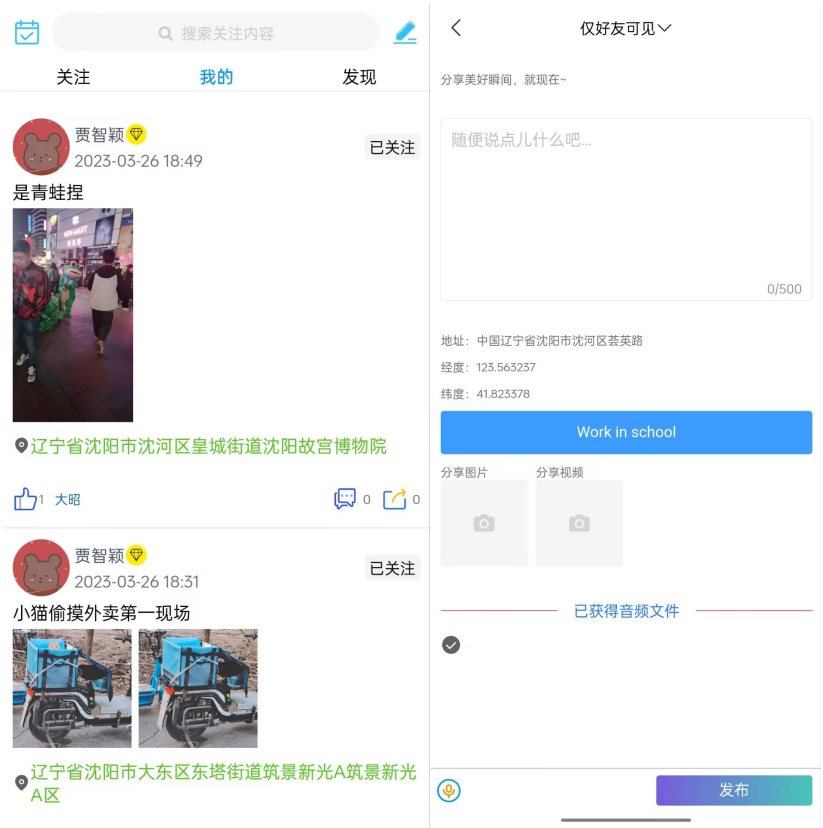

Figure 1. Lifelog APP interface
图 1. Lifelog APP 界面

APP 会根据用户当前位置通过百度 API 自动记录用户的地里位置信息。每一条 Lifelog 数据记录了基本的位置信息(经度、纬度)、个人活动信息(行为描述)和显示信息(图像或视频)。为了更好将文本数据用于科研，当用户上传行为描述文本时，系统会自动将中文转换为英文。

用户不会在特定的时间发布动态，因此 Lifelog 数据集是典型的非连续的，包含标注等丰富信息的数据集。现在任何人都可以在我们的网站(www.lifelog.vip)上免费的获取已经公开的 Lifelog 数据，也可以免费下载这个 App，从而参加到我们的项目中。图 2 展示了我们 Web 页面端收集的数据集。

num	photo	lon	lat	Description	description(English)	address	behavor	time
1		123.44972	41.71482	吃个饭，健康真好	Have a meal. It's good to be healthy	Shenyang City, my location	Work in school	2023-03-24 17:47:27
2		123.515981	41.743521	无人机加扫描搞完了，冻死了。	The drone scanning is finished and it's frozen to death.	Shenyang University of Architecture, No. 25, Hunnan Middle Road, Hunnan District, Shenyang City, Liaoning Province	Work in school	2023-03-24 17:37:54
3		123.481188	41.746027	送完Jasper上英语课。准备去学校吃饺子。	After sending Jasper to English class. I'm going to school to eat dumplings.	Wusan Street Langyue Street Linli Park, Hunnan District, Shenyang City, Liaoning Province	Chores	2023-03-24 17:18:11
4		123.51164	41.7437	可以了，成功了，上午端口没改，一直没出来	Okay, I succeeded. The port hasn't been changed this morning, and it hasn't come out yet	Shenyang Architectural University, Hall A1, Wusan Street Architectural University, Hunnan District, Shenyang City, Liaoning Province	Work in school	2023-03-24 16:33:10
5		114.769131	35.290752	真的是无语了，这字不会是知了爬的吧，眼镜快瞎了，还有一大半呢，我的卡兹兰大眼睛注定要无了。	"Really speechless, this word can't be a cicada crawling, right? My glasses are almost blind, and there is still more than half of them. My Kazilan's big eyes are destined to disappear.".	Ding Luan Zhen Hope Avenue, Chengyuan City, Xinxiang City, Henan Province	Work in school	2023-03-24 15:48:09

Figure 2. Lifelog dataset
图 2. Lifelog 数据集

2.2. 数据集

本文选取 Lifelog 数据集中用户 Liu 在 2011 年至 2022 年间发表的 9000 余条 Lifelog 数据作为数据源，在后文中我们称其为 Liu Lifelog 数据。表 1 展示了 Liu Lifelog 中的部分数据。其中 Num 为每一条数据的序号，Lon 和 Lat 分别为用户发送动态时所处地理位置的经度和纬度信息，Description 为用户所发动态的文本描述，Behavior 为动态的所属类别，Time 表示发送动态的时间。

Table 1. Liu lifelog data
表 1. Liu Lifelog 数据

Num	Lon	Lat	Description	Behavior	Time
9485	123.517748	41.744032	Rest at school	Rest	2022/2/27 12:32:45
6015	123.480839	41.696514	International Software Park meeting	Work outside	2019/4/18 12:28:17
5053	123.445885	41.737401	Eating in Hunnan	Eating outside	2018/7/18 18:12:30
6054	123.400992	41.748622	On the way to the studio	On road	2019/10/25 14:40:42

收集到的 Liu Lifelog 数据中可能包含噪音,或是存在不一致、不完整等问题，直接进行训练可能对建模效果产生不良影响。因此，必须对数据进行预处理从而提高数据的准确性，以确保在模型训练过程中更好地捕捉和理解文本信息。

- 删除空值：用户在某些时刻选择只上传图片、像音频类的文件，而未进行任何文本描述，这些记录的 Description 文本字段为空，因此我们选择剔除掉这些文本数据为空的记录。
- 删除标点符号以及特殊字符：在文本中，标点符号以及包含的一些特殊字符，如 HTML 标签，表情符号等，这些符号对于文本分类来说没有实际意义，需要将它们从文本中删去。
- 大小写转换：确保文本数据中的字母大小写是一致的，有助于消除由于大小写不一致而可能导致的混淆和误解。在文本数据中，同一个词可能以不同的大小写形式出现，例如“Rest”和“rest”。如果

不进行大小写转换，模型可能会将它们视为不同的特征。因此，我们将 Liu Lifelog 中所有文本数据转换为小写形式，从而消除由于大小写不一致而可能导致的混淆和误解。

● 去除停用词：在文本中经常出现没有实际意义，但是出现频率较高的词语，例如“a”、“the”等。这些词语不仅会增大计算量，还可能会影响模型的效果，因此，需要将他们从数据中剔除。

经过上述对数据的预处理后，我们过滤掉了 Liu Lifelog 中缺失或不完整的数据。同时由于 other 类型的数据表意不明，我们选择将这些具有干扰性的数据去除。最终，我们得到了 8 个种类的 Lifelog 数据，即 ork in school、work outside、chores、rest、on road、meal outside、entertainment、tourism。

3. DTC-TextCNN 模型

3.1. DTC-TextCNN 模型框架

为了解决 Lifelog 领域中数据分类问题，我们提出了 DTC-TextCNN 文本分类模型。模型对用户发布的文本描述进行主题特征提取，对经度和纬度信息进行聚类，将不同的地理位置分配到不同的簇中进而得到地理位置簇特征。最后，将用户发布的文本描述、文本的主题特征和地理位置特征拼接，输入到 TextCNN 中进行分类。图 3 为 DTC-TextCNN 模型的流程图。

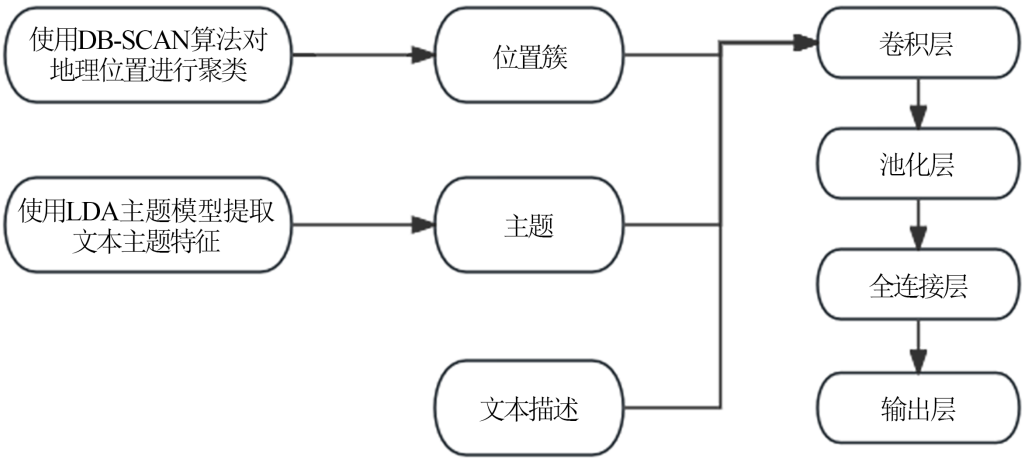

Figure 3. DTC-TextCNN model flowchart
图 3. DTC-TextCNN 模型流程图

3.2. DTC-TextCNN 模型特征提取

用户在相同地点发布的动态内容可能呈现较高的相似性，因此考虑将地理位置作为一个重要特征纳入模型，以增强对动态内容的分类能力。同时由于用户在同一个地点所发出的动态包含的经度纬度信息可能具有一定的偏差，因此我们使用 DB-SCAN 算法[16]对其进行聚类，以便识别相同的地理位置。

LDA 主题模型[17]是一种可以挖掘文档数据集合中潜在主题信息的概率模型。首先，使用 LDA 模型对 Liu Lifelog 进行训练，以发现其中的潜在主题。每一条 Lifelog 数据都被表示为一个主题分布，其中每个主题与一组词相关联。利用 LDA 生成的主题分布，将文本转换为一个主题向量。这个向量可以被视为文本在主题空间中的表示，它捕捉了文本中潜在的语义信息。

人类语言具有高度模糊性，一句话可能有多重的意思或隐喻，而计算机当前还无法真正理解语言或文字的意义。因此，现阶段的主要做法是将文本数据转换为模型可以理解的数值表示，这通常包括词嵌入[18]的应用，将每个词语映射到一个实数向量。Word2Vec [19]是一种用于将词语表示为向量的技术，

能够捕捉词语之间的语义关系[20]。通过 Word2Vec，每个单词都被映射为一个固定维度的向量，这些向量保留了单词之间的语义关系，使得它们可以被用作神经网络等深度学习模型的输入。

使用 LDA 模型训练的主题向量和 Word2Vec 方法训练的词向量作为文本分类器的输入。Word2Vec 侧重于词语上下文信息，但在表达词语的全局语义时存在不足。相比之下，LDA 模型训练的主题特征向量反映了词语在整个文档集中的全局主题特征，更好地概括了词语的全局语义。因此，我们在文本分类时引入了 LDA 模型中词语的主题特征作为额外的语义补充信息，以提升分类效果。

3.3. DTC-TextCNN 模型搭建

由于在 Liu Lifelog 中，所记录的数据通常都是短文本数据，即这些记录包含相对较短的文本片段，可能是几个句子、一个段落，或者是简短的描述。因此采用短文本分类的方法能更好地对 Liu Lifelog 数据进行分类。CNN 相比于其他的深度学习模型以及经典的分类方法，对于文本的局部信息比较敏感；计算开销低。同时，在对短文本进行分类中，TextCNN 是比较好的处理模型,在进行实验验证时具有较好的分类效果[21]。

通过 DB-SCAN 聚类算法进行地理位置的特征提取，在本文中，使用球面距离来衡量两者之间的距离并作为聚合的半径参数。选取 1 公里作为密度聚合的半径参数，MinPts 的个数为 5。聚类后部分结果如表 2 所示，其中 Cluster 字段为聚类后的位置簇 id 号。

Table 2. Geographical location clustering results
表 2. 地理位置聚类结果

Description	Lon	Lat	Cluster
Rest at school	123.517748	41.744032	80
Eat with students	123.509232	41.742581	80
International Software Park meeting	123.480839	41.696514	351
Eating in Hunnan	123.445885	41.737401	4

文本主题是由 3.2 节中提到的 LDA 主题模型训练得到的，使用 LDA 主题模型，生成文档－主题分布以及主题－词语分布。在本文中，由于需要进行分类的数据共有 8 种，因此我们将模型的主题参数 k 值设置为 8。经过训练后生成的“主题－词语”分布可以得到每个词语关于主题的概率值，它表示每个词在主题下的概率分布，这可以更加丰富地表示文章中每一个词的潜在语义，以获得更加准确的效果。一个词在主题中的概率越大，说明这个词在这一主题中的重要程度越高，也就越能够表征该主题的主题信息，也就应赋予该词更高的权重。我们选取每个主题下概率前十的词语，计算在其主题下所占的归一化权重，公式如(1)所示。其中 $p\left(w_i \middle| z_t\right)$ 表示在第 t 个主题下词语 i 出现的概率。将通过(1)式计算得到的每个词在相应主题下的归一化权重与通过 Word2Vec 方法训练获得的词向量通过加权求和得到相应的主题向量，公式如(2)所示。其中 V_{wi} 为通过 Word2Vec 方法训练获得的词向量，V_{topic_t} 为每一段生活日志文本对应的主题向量。主题向量的获取如流程如图 4 所示。

$$\theta_{W_{t,i}} = \frac{p\left(w_i \middle| z_t\right)}{\sum_j^{10} p\left(w_i \middle| z_t\right)} \tag{1}$$

$$V_{\text{topic}_t} = \sum_i^K \theta_{W_{t,i}} \times V_{wi} \tag{2}$$

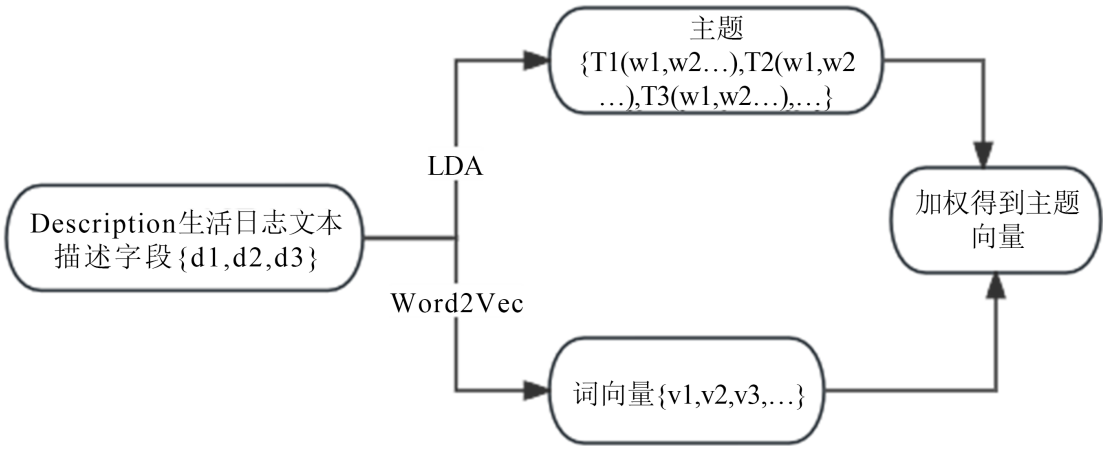

Figure 4. Topic Vector Generation Process
图 4. 主题向量生成过程

将向量化后的文本信息，位置信息同主题向量拼接，使用两个不同卷积核大小的卷积层，然后通过全局最大池化层提取最显著的特征，最后通过全连接层进行分类。表 3 详细描述了 DTC-TextCNN 模型的整体结构，其中包括各层的参数数量和输出形状。

Table 3. DTC-TextCNN model structure
表 3. DTC-TextCNN 模型结构

Layer (type)	Output Shape	Param
input_1 (InputLayer)	(None, 120)	0
embedding_1 (Embedding)	(None, 120, 150)	762,300
conv1d_1 (Conv1D)	(None, 120, 512)	230,912
conv1d_2 (Conv1D)	(None, 120, 128)	57,728
global_max_pooling1d_1(GlobalMaxPooling)	(None, 512)	0
global_max_pooling1d_2(GlobalMaxPooling)	(None, 128)	0
concatenate_1 (Concatenate)	(None, 640)	0
dropout_1 (Dropout)	(None, 640)	0
batch_normalization_1 (BatchNormalization)	(None, 640)	2560
dropout_2 (Dropout)	(None, 640)	0
dropout_3 (Dropout)	(None, 640)	0
dense_1 (Dense)	(None, 6)	3864

4. 实验结果

本节使用 2.2 节介绍过的 Liu Lifelog 数据进行实验。将 Liu Lifelog 数据集按照 8:2 的比例随机划分成互不相交的两部分作为训练集和测试集。使用 Dropout 正则化技术[22]，以防止模型在训练集上的过拟合，提高其在未见过的数据上的性能。使用批归一化加速神经网络训练过程，缓解梯度消失问题，使得训练更加稳定[23]。表 4 展示了 DTC-TextCNN 模型的实验环境。

Table 4. DTC-TextCNN model experimental environment
表 4. DTC-TextCNN 模型实验环境

Factor	Contents
Activation Function	Relu
Cost Function	Categorical_crossentrop
Learning rate	Auto
Optimizer	Adam
Epochs	30

使用 DTC-TextCNN 对 Liu Lifelog 数据集分类的总体准确率为 60.9%。此外，我们还建立了只包含地理位置信息和动态文本信息的 DC-TextCNN 模型，以及只包含动态文本信息的 D-TextCNN 模型，这两个模型对于数据集分类的准确率分别为 57.9%和 50.1%。上述模型的验结果如表 5 所示。可以看到，融合了主题特征和地理位置特征的 DTC-TextCNN 能够有效提高生活数据的分类效果。

Table 5. Experimental results of DTC-TextCNN model
表 5. DTC-TextCNN 模型实验结果

Model	Accuary
DTC-TextCNN	60.9%
DC-TextCNN	57.9%
D-TextCNN	50.1%

5. 结论

对 Lifelog 数据的分类不同于传统的文本分类，因为它是专门记录日常生活的经历，与个体用户的生活和经验相关联，涉及用户的兴趣、活动、位置等信息，这些个性化的特点要求用于分类的模型需要更好地适应用户的特殊行为和偏好。在本文中，我们提出了 DTC-TextCNN 模型用于解决 Lifelog 领域中数据分类问题。该模型融合了地理位置信息和文本主题信息。相比于基本模型，该模型能够提高 Lifelog 数据分类的准确性。

对 Lifelog 数据的收集是一件有意义的事情，人们可以通过对以往数据的查看，来回忆发生在自己身上的那些值得纪念的事情，同时，对 Lifelog 进行分类，使得人们更加方便地搜索相关动态。现在我们的项目已经公开在 www.lifelog.vip，同时提供相关数据集供公众研究使用。希望有更多的志愿者能参与到项目之中，和我们一起共同留存生活中的美好时刻。

参考文献

[1] Garcia, F.C.C., Hirao, A., Tajika, A., Furukawa, T.A., Ikeda, K. and Yoshimoto, J. (2021) Leveraging Longitudinal Lifelog Data Using Survival Models for Predicting Risk of Relapse among Patients with Depression in Remission. 2021 43*rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (*EMBC*), Mexico, 1-5 November 2021, 2455-2458. https://doi.org/10.1109/EMBC46164.2021.9629798

[2] Gurrin, C., Smeaton, A.F. and Doherty, A.R. (2014) LifeLogging: Personal Big Data. *Foundations and Trends® in Information Retrieval*, **8**, 1-125. https://doi.org/10.1561/1500000033

[3] Dobbins, C., Rawassizadeh, R. and Momeni, E. (2017) Detecting Physical Activity within Lifelogs towards Preventing Obesity and Aiding Ambient Assisted Living. *Neurocomputing*, **230**, 110-132.

[4] Choi, J., Choi, C., Ko, H., *et al*. (2016) Intelligent Healthcare Service Using Health Lifelog Analysis. *Journal of Med-*

ical Systems, **40**, 1-10. https://doi.org/10.1007/s10916-016-0534-1

[5] Li, D.L., Gu, Y. and Kamijo, S. (2018) Smartphone Based Lifelog with Meaningful Place Detection. 2018 *IEEE International Conference on Consumer Electronics* (*ICCE*), Las Vegas, 12-14 January 2018. 1-5. https://doi.org/10.1109/ICCE.2018.8326116

[6] Liu, G., Rehman, M.U. and Wu, Y. (2021) Personal Trajectory Analysis Based on Informative Lifelogging. *Multimedia Tools and Applications*, **80**, 22177-22191. https://doi.org/10.1007/s11042-021-10755-w

[7] Liu, G.Q., *et al*. (2014) Behavior Identification Based on Geotagged Photo Data Set. *The Scientific World Journal*, **2014**, Article ID: 616030. https://doi.org/10.1155/2014/616030

[8] Liu, G., Rehman, M.U. and Wu, Y. (2021) Toward Storytelling from Personal Informative Lifelogging. *Multimedia Tools and Applications*, **80**, 19649-19673. https://doi.org/10.1007/s11042-020-10453-z

[9] Maron, M.E. (1961) Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, **8**, 404-417. https://doi.org/10.1145/321075.321084

[10] 马新宇, 黄春梅, 姜春茂. 基于三支决策的 KNN 渐进式文本分类方法[J]. 计算机应用研究, 2023, 40(4): 1065-1069.

[11] Kalcheva, N., Karova, M. and Penev, I. (2020) Comparison of the Accuracy of SVM Kemel Functions in Text Classification. 2020 *International Conference on Biomedical Innovations and Applications* (*BIA*), Varna, 24-27 September 2020, 141-145. https://doi.org/10.1109/BIA50171.2020.9244278

[12] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. https://doi.org/10.3115/v1/D14-1181

[13] 王佳慧. 基于 CNN 与 Bi-LSTM 混合模型的中文文本分类方法[J]. 软件导刊, 2022, 1(22): 159-163.

[14] 杨阳, 刘恩博, 顾春华, 等. 稀疏数据下结合词向量的短文本分类模型研究[J]. 计算机应用研究, 2022, 39(3): 711-715, 750.

[15] Sharma, A.K., Chaurasia, S. and Srivastava, D.K. (2020) Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec. *Procedia Computer Science*, **167**, 1139-1147. https://doi.org/10.1016/j.procs.2020.03.416

[16] Schubert, E., Sander, J., Ester, M., *et al*. (2017) DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems* (*TODS*), **42**, 1-21. https://doi.org/10.1145/3068335

[17] Jelodar, H., Wang, Y., Yuan, C., *et al*. (2019) Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimedia Tools and Applications*, **78**, 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

[18] Bakarov, A. (2018) A Survey of Word Embeddings Evaluation Methods.

[19] Mikolov, T., Chen, K., Corrado, G., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781

[20] Ma, L. and Zhang, Y. (2015) Using Word2Vec to Process Big Text Data. 2015 *IEEE International Conference on Big Data (Big Data)*, Santa Clara, 29 October 2015 - 1 November 2015, 2895-2897. https://doi.org/10.1109/BigData.2015.7364114

[21] Zhang, T. and You, F. (2021) Research on Short Text Classification Based on Textcnn. Journal of Physics: Conference Series. *IOP Publishing*, **1757**, Article ID: 012092. https://doi.org/10.1088/1742-6596/1757/1/012092

[22] Srivastava, N., Hinton, G., Krizhevsky, A., *et al*. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, **15**, 1929-1958.

[23] Ioffe, S. and Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the* 32*nd International Conference on International Conference on Machine Learning*, **37**, 448-456.