基于 BERT 深度语言模型的“智慧政务”文本挖掘应用

目录

摘要： 本文基于近年来语言智能的里程碑技术：BERT 深度语言模型，围绕“智慧政务”文本挖掘的主题，完成了（1）留言自动分类、（2）热点问题挖掘、（3）答复意见质量评价 3 项工作。在任务（1）上，BERT 模型在测试集上的 F-score 达到了 92.89%，明显优于基线模型 LSTM 与 Fasttext。在任务（2）上，我们创新提出了基于 Affinity Propagation 的算法对留言进行初步聚类，再通过命名实体识别进一步过滤异常值的策略。并结合话题时长、留言数、留言的点赞与反对数共 3 类因素，综合评价了每个问题的热度值。在任务（3）上，我们从相关性、可解释性、完整性等角度出发，提出了 9 项特征，综合评价答复意见。这有助于客观、全面反映答复意见的质量，提高政府工作人员的办事水平与群众的满意度。

关键词：BERT 深度模型；智慧政务；文本分类；AP 聚类；命名实体识别；热点挖掘；答复质量评价

Abstract: Based on the milestone technology in language intelligence: BERT, focused on the topic “government administration Intelligence”, this paper completed 3 tasks: (1)automatic message classification, (2) hot spots mining, (3) evaluating the quality of reply. For task (1), the F-score of BERT on test set reached 92.89%, which is superior to baseline LSTM and Fasttext model. For task(2), we put forward a new strategy which firstly did message clustering using Affinity Propagation algorithm, then further excluded outliers through Named Entity Recognition. We combined 3 features: duration, the number of messages, and the number of “like” and “dislike” votes, thus evaluating the degree of heat of every topic. For task(3), taking correlation, interpretability and integrity into consideration, we put forward 9 features to assess each reply. This system is helpful to reflect the quality of each reply objectively and completely, and improve the working level of civil servants and the satisfaction of people.

Key words: BERT deep model; government administration Intelligence; text classification; AP clustering; Named Entity Recognition; hot spots mining; the evaluation of reply

1. 绪论

1.1 “智慧政务”文本挖掘的意义

近年来，随着“互联网+政务”服务的推进，市长信箱、民意留言板、阳光热线等网络问政平台逐步成为政府已经成为政府了解民情、听取民声、体察民意、汇聚民智的一个重要桥梁。同时，随着大数据、云计算、人工智能特别是语言智能等技术的不断突破，建立起基于自然语言处理（Natural Language Processing, NLP）技术的智慧政务系统，已成为社会治理创新发展的迫切需求与新趋势。如何运用NLP 技术，批量、智能、高效地处理海量的政务文本，进而建立智能化的电子政务系统，是服务型政府建设中的一个重要子课题。这对于提升政府的施政效率与治理水平，增强人民群众的幸福感，促进社会和谐，都具有重大的积极意义。

1.2 “智慧政务”文本挖掘的目标

“智慧政务”文本挖掘的目标主要包含 3 部分，分别是（1）群众留言的自动分类；（2）群众留言的热点话题发现；（3）留言答复意见的质量评价。

（1） 群众留言的自动分类

许多网络问政平台，每天都会接收大量的群众留言。平台的工作人员首先按照预先设置的分类体系，对留言进行归类。这便于将数目浩繁的留言分派至相应的职能部门处理，对症下药。目前，大部分政务系统的群众留言，还依赖于人工凭直觉分类。不仅工作量大、效率低，而且差错率高。因此，利用自然语言处理中的文本分类（Text classification）技术实现留言自动分类，能极大地减轻政务工作人员的负担。

（2） 群众留言的热点话题发现

在海量的群众留言中，存在着许多反映共同问题、表达共同诉求的留言。对它们进行针对性地处理，有利于分清民情诉求的轻重缓急，提升政府服务的质量与效率。这属于 NLP 中的话题检测与跟踪（Topic Detection and Tracking，TDT）的课题范畴。因此，我们需要探索如何从大量留言中，自动发现某一时段内群众集中反映的热点问题。

（3） 留言答复意见的质量评价

对于每一条群众留言，政府工作人员会对其答复，回应问题的处理情况，告知相关政策规定，或提供建议意见等。自动地评价答复意见的质量，有助于将群众的诉求落到实处，改善政府的办事水平。因此，在论文的第四章，我们将融合 9 项指标，从相关性、完整性、可解释性、条理性等角度，自动地综合评价答复意见的质量。

1.3 语言智能的里程碑技术：BERT 深度语言模型介绍

基于神经网络架构的深度学习算法由于其能自动提取数据特征，以及其强大的拟合泛化能力，已经在计算机视觉（CV）、自然语言处理（NLP）、机器人（robotics）、推荐系统（recommendation system）等多个人工智能领域取得了重大突破。自从 2013 年谷歌的 Mikolov 团队提出词汇语义表示模型 word2vec[1]后，海量文本中的每一个词都被表示为一个稠密、低维的实值向量，自然语言处理领域也进入了深度学习时代。

近年来，NLP 界以 ELMo[2]、BERT[3]代表的预训练深度语言模型（Pre-trained Language Model）在以往神经网络模型的基础上，进一步改善了文本语义表示的效果，并在文本分类、命名实体识别、信息抽取、人机对话、机器翻译、阅读理解等 NLP 各项下游任务中取得了重大突破，频繁且大幅度刷新了之前地最好结果。例如 2018 年 Google 团队发布的 BERT 模型，在 11 项不同的 NLP 测试中，均表现出最佳效果，将通用语言理解评估（GLUE）基准提升至 80.4%，超出以往最佳模型 7.6%[4]。以 BERT 为代表的深度语言模型已经成为 NLP 里程碑式的技术。

预训练深度模型应用于下游任务，主要分为两种策略：

一是基于特征的（feature based）策略，即固定的语言特征向量从模型中提取出来服务于后续任务，以 ELMo 模型为代表。

二是微调（Fine-tuning）策略，即在模型顶部添加着眼于具体任务的分类层，并且模型所有的参数也随着下游任务的训练适度优化。微调策略实质上是一种迁移学习（Transfer Learning），可以充分利用已训练的深度模型，迁移到新的任务上。与从零开始训练模型相比，微调不仅节省了大量的计算开销，也显著提高了

模型的精度。而 BERT 模型就是采用微调策略的预训练模型的代表[3]。

BERT (Bidirectional Encoder Representations from Transformers）是一种基于 Transformer 架构的预训练深度学习语言模型，其结构主要如图 1 所示：

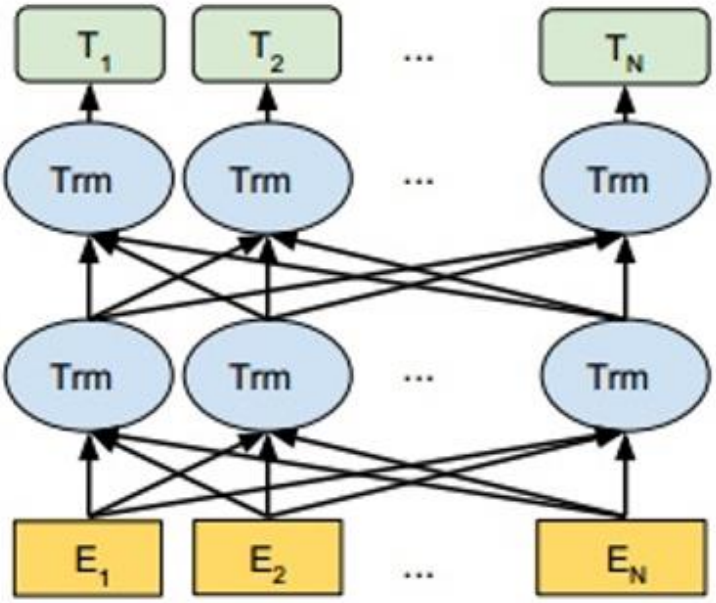

图 1 BERT 模型基本架构

以中文的 BERT 预训练模型为例，图 1 的 E1,E2,…EN 表示在首尾分别添加[CLS]和[SEP]标记的中文字符。它们依次经过 12 或 24 层双向的 Transformer（Trm）编码器，就可以得到文本字符语境化的向量表示（Contextual Embedding）。Transformer 是一个基于自注意力（self-attention）机制的编码-解码器[5]。最底层的 Transformer 编码器的输入为字符向量、字符位置向量与句子片段向量之和。模型内每一层均由多头自注意力(Multi-head Self-attention) 和前馈神经网络（Feed-forward Neural Networks）两部分构成，前者使编码器在给每个字符编码时，能关注到周围其他字符的信息；后者用于增强模型的拟合能力。模型的每一层经过一个相加与归一化（add & norm）操作后，生成新的字符向量，作为下一层编码器的输入。顶层编码器输出的[CLS]标记的向量 T1，可以视为整个句子的语义表征；而顶层编码器输出的向量 T2,T3……则分别是字符 E2,E3 语境化的向量表示。它们为文本分类、命名实体识别等后续任务提供了重要支撑。

另外，为增强语义表示的能力，BERT 提出了两个预训练的目标任务：遮罩语言模型（Masked LM, MLM）和下句预测（Next Sentence Prediction, NSP）。MLM 实质是一个完型填空任务，中文语料中 15%的字会被选中，其中的 80%被替换为

[MASK]，10%被随机替换为另一个字，剩下的 10%保持原字。模型需要根据句中的其他字，生成被选中字位置的向量，经由一个线性分类器，预测被选中的字。被选中的 15%的字之所以没有全部替换为[MASK]，是因为该遮罩标记在实际下游任务的语料中并不存在。出于与后面任务保持一致的考虑，BERT 需按一定的比例在预测的字的位置放置原字或者某个随机字，使得模型更偏向于利用上下文信息预测被选中字。在下句预测任务中，模型选择若干句子对，其中有 50%的概率两句相邻，50%的概率两句不相邻。模型通过上述两个目标任务，能够较好地学习到文本中字词和句间的语义信息。

考虑到 BERT 模型在语言智能领域的显著优势，本文拟将该模型运用到“智慧政务”文本挖掘之中。采取微调（Fine-tuning）的迁移学习策略，在文本分类、文本聚类、命名实体识别（Named Entity Recognition, NER）等具体任务上，充分发挥 BERT 中文模型[1]的功效。

1.4 本文的总体框架

本文的总体组织框架如下：

第1章　绪论

介绍“智慧政务”文本挖掘的意义和总体目标；介绍本文主要采用的深度语言模型 BERT 的原理；介绍本文的总体框架与主要创新点。

第2章　基于 BERT 模型的留言自动分类

首先，介绍该任务的目标与实验数据集；第二，设计基于 BERT 模型的留言自动分类算法；最后，对比分析 BERT 模型、LSTM 模型、FastText 模型在文本分类上的效果（各类的 F1 值与整体的 F-score）。

第3章　基于语义相似度与 BERT 命名实体识别的留言热点问题挖掘

首先，介绍该任务的目标与实验数据集；第二，基于 BERT 模型，将每条留言的语义向量与命名实体[2]（Named Entity）作为特征表示；第三，利用基于图的 Affinity Propagation（AP）聚类算法，实现留言的无监督自动聚类；

[1] 原始的 BERT 中文模型，由谷歌公司在海量的中文维基百科数据上训练而成。
[2] 命名实体（Named Entity）指文本中的人名、地名、机构名、时间等专有名词。命名实体识别（NER）是自然语言处理领域的一项重要任务。

最后，对于每个聚类后的话题（问题），综合考虑（1）话题时长、（2）话题包含的留言数量；（3）每条留言的点赞数与反对数 3 项指标，设计话题热度评价体系并予以实现。

第4章　多特征融合的答复意见质量评价

首先，介绍该任务的目标与实验数据集；第二，根据答复意见的长度、答复的及时性、关键词覆盖率、答复与留言的相关性、答复的专业性等 9 项特征指标，综合评价答复意见的质量。最后，在程序上实现质量评价系统，并进行效果分析。

第5章　结语

总结本文的工作，展望今后的改进方向。

1.5 本文主要的创新之处

（1）将语言智能领域最新的 BERT 深度模型应用于政务文本挖掘。基于 BERT 的留言分类模型明显优于前人的 LSTM 模型与 Fasttext 模型。

（2）针对以往热点话题聚类模型健壮性（Robustness）差、效果不佳的问题，提出了利用命名实体识别（NER）任务增强留言区分度的策略，进而显著改善了留言无监督聚类的效果。

（3）针对以往无监督聚类任务不知如何预设聚类数量的问题，采用了基于距离的 Affinity Propagation（AP，亲和力传播）的聚类算法。使得自动确定聚类数量成为可能，节省了大量的试错成本。

（4）从内容丰富度、答复相关度、答复专业性、答复时效性等角度，提出了“9 项合一”答复意见质量的评测方法。更加全面地反映了政府工作人员反馈群众留言的水平。

2. 基于 BERT 模型的留言自动分类

2.1 任务介绍与实验数据集

该任务属于自然语言处理中的文本分类任务。附件 2 包含了 9210 条群众在网

络平台上发布的留言，分为城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生共 7 类。每条留言均包含留言主题、留言详情的字段。我们将数据集打乱顺序后，按照 8:1:1 的比例，分别划分训练集、验证集和测试集。包括 BERT 在内的所有模型，均在相同的训练集、验证集上进行训练、调整超参数，并在相同的测试集上进行测试。

2.2 实验流程

在输入的文本方面，考虑到 BERT 模型允许单一文本的最长长度为 512 个字符，加之有 Sun（2019）[6]等人的相关实验表明，长文本截取前 512 个字符，已能在 BERT 模型中取得理想的分类效果，我们拼接了每一条留言的主题文本与详情文本，截取前 512 个字符作为模型的输入。

在超参数设置方面，我们参考 Sun（2019）[6]等人在 BERT 上的文本分类经验，如下设置超参数：学习率 lr=2e-5，衰变因子 ξ=0.95。此外，训练遵循早停（early stopping）原则，当模型的损失在验证集上不再下降，就视为模型在验证集上已经收敛，可以停止训练。这能够有效地避免过拟合（Overfitting）问题，保证模型的泛化能力以及在测试集上的表现。

如 1.3 节所述，对于文本分类任务，BERT 模型提取顶层的符号[CLS]的特征向量 v（768 维），作为整个文本的特征表示，再后接一个 768*n 的全连接层（Fully-connected layer）W（n 为文本类别数），最后通过 softmax 函数归一化，输出一个文本分别属于各个类别 c 的概率：

$$P(c|v) = \text{softmax}(W \cdot v)$$

其中 softmax 函数：

$$\text{softmax}(x_c) = \frac{\exp(x_c)}{\sum_{i=1}^{n} \exp(x_i)}$$

在训练过程中，模型会调整全连接层 W 以及 BERT 12 层模型的参数，使得每个文本的正确类别所对应的概率最大化。

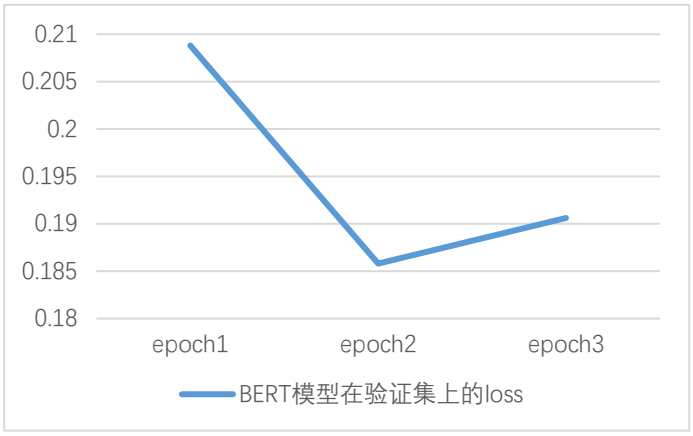

图 2　BERT 模型在验证集上的损失变化图

如图 2 所示，当 BERT 模型在训练第 3 轮（epoch）时，在验证集上的损失开始上升。

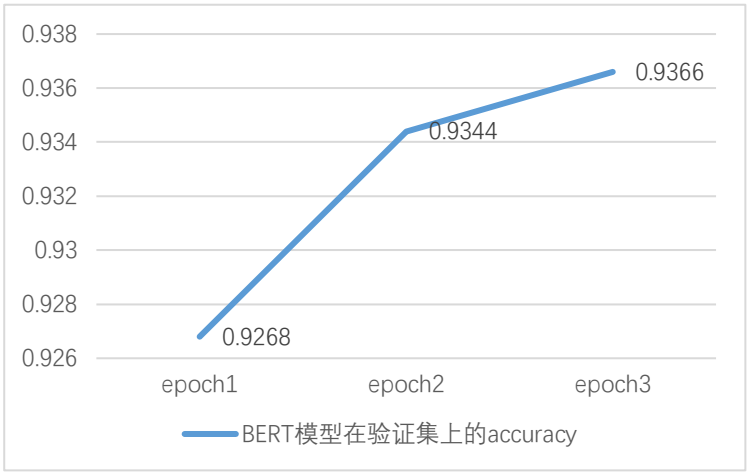

图 3 BERT 模型在验证集上的 accuracy 随训练轮数的变化情况

图 3 关于 BERT 模型在验证集上的正确率（accuracy）变化情况也表明，在第 2 轮训练时，分类的正确率较前一次明显提升约 0.76%；但第 3 轮训练的 accuracy 提升已不明显。因此，综合模型在验证集上的损失与正确率，根据早停（early stopping）原则，我们认为第 2 轮训练结束后的 BERT 分类模型已经收敛，可以作为最终模型。

2.3 BERT 分类效果及其与其他模型的对比分析

最终收敛的 BERT 模型在测试集上测试的结果如表 1 所示：

留言类别	查准率 P	查全率 R	F1 值
城乡建设	95.68%	88.50%	91.95%
卫生计生	89.77%	96.34%	92.94%
商贸旅游	88.19%	91.80%	89.96%
劳动和社会保障	**96.89%**	94.92%	95.90%
教育文体	95.63%	**96.84%**	96.23%
交通运输	82.61%	91.94%	87.02%
环境保护	96.77%	95.74%	**96.26%**

表 1: BERT 模型在测试集上的各类分类效果

为证明 BERT 模型在留言分类任务上的良好能力，我们将基于字向量的 LSTM(Long Short-Term Memory)[7]和基于 Fasttext[8]的 2 个文本分类模型作为基线（baseline）模型，比较 BERT 与基线模型的效果。

LSTM 模型作为循环神经网络（RNN）的变种，是一种基于时间序列的链式结构。它克服了传统 RNN 模型梯度消失的缺陷，成为近年来 NLP 领域应用较广泛的特征提取器。我们使用的 LSTM 模型为加入了 dropout 机制的通用改良版本[9]。dropout 机制能有效避免 LSTM 模型的过拟合问题。与基于 BERT 的文本分类模型类似，输入的文本经过 LSTM 隐层后，通过 softmax 归一化层，对 LSTM 隐层传递来的信息进行学习, 并计算出待分类文本属于各类别的概率[10]。

Fasttext 模型是 2017 年 Facebook 公司 AI 团队提出的一种快速文本分类的模型。其基本架构如图 4 所示，它包含 3 个部分：input layer 输入层、hidden layer 隐藏层和 output layer 输出层。首先；输入层的 $x_1,x_2,\ldots,x_{N-1},x_N$ 表示一个文本中的 N 个 n-gram 向量。其次，隐藏层将向量特征求和取平均，并采用单层神经网络学习。最后在输出层，通过一个线性分类器，输出一个文本分别属于各类别的概率[8]。

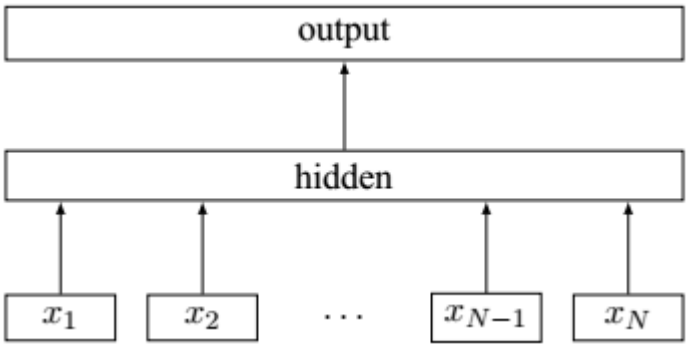

图 4 Fasttext 文本分类模型的基本结构

LSTM 文本分类模型	Fasttext 文本分类模型
隐藏层节点数：128	字向量维度：100 维
优化算法：Adam	学习率：0.1
batch_size：4	增强语义的 N-gram 类型：2-gram

表 2: LSTM 与 Fasttext 模型的超参数设置情况

LSTM 和 Fasttext 模型超参数设置情况如表 2 所示。BERT 模型与上述两个基线模型均使用相同的训练、验证、测试集。

文本分类通用评价指标为查准率（Precision, P）、查全率（Recall, R）与 F1 值。如表 3 所示，各模型在测试集上预测了每条留言所属的类别后，可以得到一个分类结果的混淆矩阵：

预测的类别 / 真实的类别	预测为正类	预测为负类
真实为正类	TP(真正类)	FN(假负类)
真实为负类	FP(假正类)	TN(真负类)

表 3 分类结果混淆矩阵

已知每一类的混淆矩阵后，该类的查准率、查全率如下两式计算：

$$\text{查准率 P}=\frac{TP}{TP+FP}$$

$$\text{查全率 R}=\frac{TP}{TP+FN}$$

F1 值综合了上述两个指标，是它们的调和均值：

$$\text{F1} = \frac{2*P*R}{P+R}$$

如果要评测一个模型在综合的分类能力，可用 F-Score 衡量：

$$\text{F-score} = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

其中 n 为类别数量，P_i是第 i 类查准率；R_i是第 i 类的查全率。

我们先观察三个模型在测试集上的 F1 值，如图 5 所示：

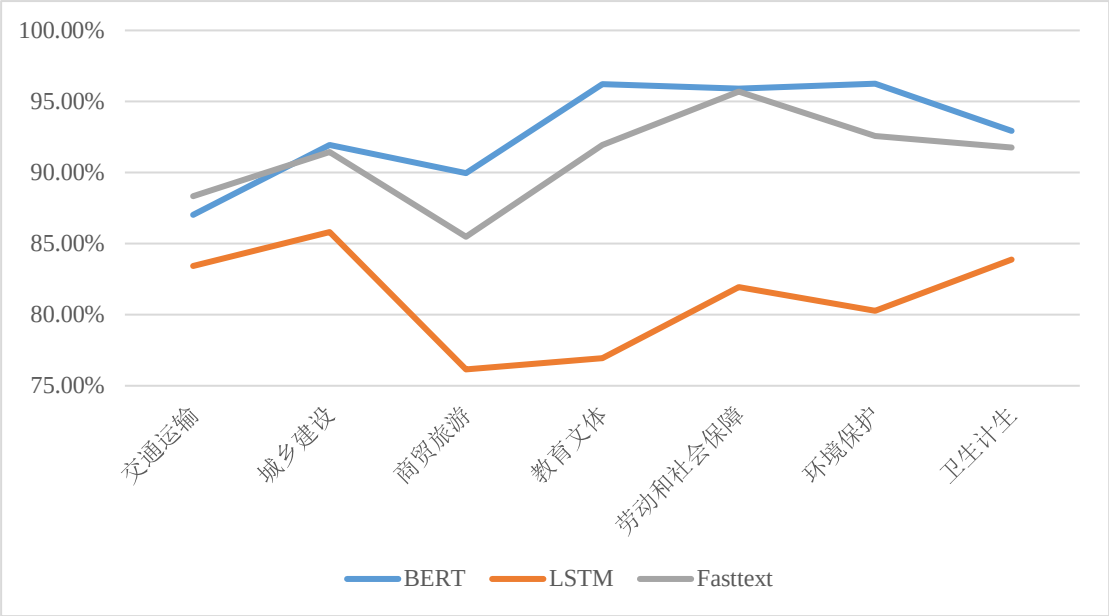

图 5 三个模型在测试集各类上的 F1 值对比

如图 5 所示，在 7 类留言文本中，BERT 在 6 类上的分类 F1 值均取得了最佳效果。仅在交通运输这一类上，Fasttext 的分类效果略优于 BERT。

衡量各模型综合分类能力的 F-score 分数如表 4 所示：

BERT	92.89%
LSTM	81.21%
Fasttext	91.03%

表 4 各分类模型的 F-score

可以看出，基于 BERT 的留言文本分类算法取得了最佳效果，其 F-score 高出 LSTM 模型约 11.6%，也高出广泛应用的文本分类模型 Fasttext 约 1.8%。BERT 深度语言模型高达 92.89%的 F-score 证明了它在政务留言文本分类任务上优越性与实用性。

3. 基于语义相似度与 BERT 命名实体识别的热点问题挖掘

3.1 任务介绍与实验数据集

从大量群众留言中及时发现热点问题，有助于政府工作分清轻重缓急，相关部门能有针对性地处理，提升服务效率。附件 3 包含了 4326 条群众的留言，每条留言均包括留言编号、留言用户、留言主题、留言时间、留言详情、反对数、点赞数一共 7 个字段。

首先，我们需要根据语义相似度，尽可能将描述同一问题的留言聚为一类。再设计合理的热度评价指标，计算每一类问题的热度值并排序。

3.2 无须预设聚类数目的 AP 聚类算法介绍

在该任务中，我们将对 4326 条群众留言进行无监督聚类。由于我们预先不知道留言大致的类别数量，如果采用诸如 K-means 等需要预设聚类数目的算法，将会消耗大量的人力与计算开销，试错成本较高。因此，经过权衡比较之后，我们采用一种无须预设聚类数目的 Affinity Propagation 聚类（简称 AP 聚类）算法。本节将概述 AP 聚类的基本原理。

AP 聚类是 2007 年 Frey 等人在著名科学杂志 *science* 上提出的算法[11]。它根据 N 个数据点之间的相似度进行聚类。AP 算法不需要事先指定聚类数目，而是将所有的数据点都作为潜在的聚类中心，称之为 exampler。

N 个数据点之间的相似度，就组成一个 N*N 的相似度矩阵 S，并以对角线上的值 S(i,i)作为第 i 个数据点能否成为聚类中心 k 的评判依据，该值越大，表明该数据点成为聚类中心的可能性也就越大。S(i,i)也称之为参考度 p（preference）。

AP 算法中传递两种类型的信息：

一是吸引度（responsibility）r(i, k)。它表示从点 i 发送到聚类中心 k 的数值信息，反映出 k 点作为 i 点聚类中心的合适程度。

二是归属度（availability）a(i, k)。它表示从聚类中心 k 发送到 i 的数值信息，反映出 i 点选择 k 点作为聚类中心的合适程度。

可以看出，吸引度和归属度越强，则 k 点作为 i 点聚类中心的可能性就越大。AP 算法就是通过多次迭代，更新每一个点的吸引度和归属度信息。当已经历最大迭代次数，或数值收敛，则对于任意点 i，计算它与所有样本的吸引度与之和，

那么 i 点的聚类中心 k 点如下式选择：

$$k = \underset{k}{\operatorname{argmax}}(a(i,k) + r(i,k))$$

如上所述，AP 聚类的核心，就是吸引度（responsibility）、归属度（availability）两个信息量交替更新的流程。吸引度 r(i,k)的迭代更新公式：

$$r(i,k) = S(i,k) - \max_{k`,k` \neq k} \{a(i,k`) + S(i,k`)\}$$

上式表示 r(i,k)由点 i 与点 k 的相似度，减去点 i 和其他点的相似度与归属度之和的最大值。

归属度 a(i,k)的迭代更新公式：

if $i \neq k$:

$$a(i,k) = \min\{0, r(k,k) + \sum_{i`, i` \notin \{i,k\}} \max\{0, r(i`,k)\}\}$$

else:

$$a(k,k) = \sum_{i`, i` \notin \{i,k\}} \max\{0, r(i`,k)\}$$

AP 聚类具有如下优势：（1）不需要事先指定聚类的数量。聚类的数量，由参考度(preference) S(i,i)的初始值与数据的分布共同决定；（2）聚类的结果不会多次运行而随机变化。这比通用的 k-means 聚类更加稳定；（3）适用于非对称与稀疏的相似性矩阵[12]。

3.3 热点问题总体挖掘流程

留言热点问题自动挖掘的总体流程如图 6 所示：

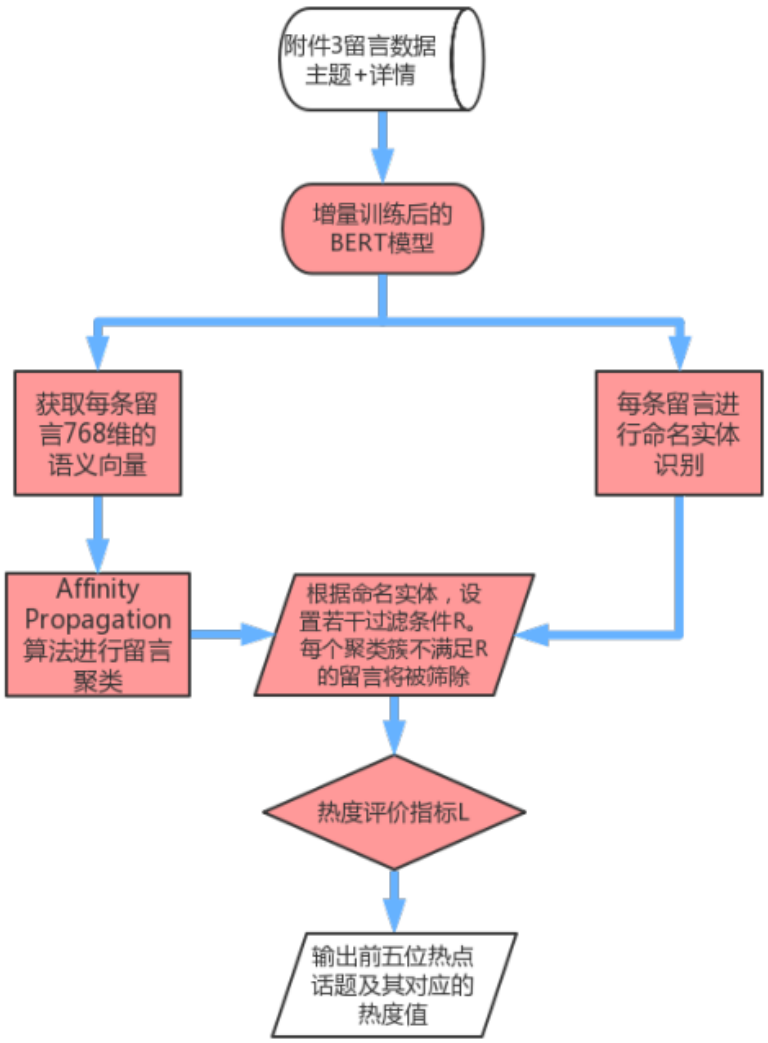

图 6 留言热点话题自动挖掘的总体流程

第一步：BERT 模型增量训练，使之学习到留言的句法语义信息

如前文 1.3 节所述，BERT 使用了两个训练任务来增强语义表示的能力：Masked Language Model 遮罩语言模型（MLM）与 Next Sentence Prediction 下句预测（NSP）。前者是一个完形填空任务，让模型根据上下文去预测被遮住的词语；后者是根据前一句话，预测下一句话的内容[3]。我们以每条留言的主题与详情拼接，作为输入文本，在 MLM 与 NSP 任务上对 BERT 模型进行增量训练。这使得 BERT 中文模型能更好地适用于我们的语料数据。

第二步：获取留言的语义向量，利用 AP 聚类算法初步聚类

利用增量训练的 BERT，获取每条留言的语义向量表示。利用无须预设聚类数目的 AP 聚类算法，将 4326 条留言初步聚为 701 类。关于 AP 聚类的基本原

理，我们在 3.2 节中已作阐述。

第三步：基于 BERT 模型，识别留言中的命名实体。设置条件集合 R，过滤异常值

由于聚类是一种无监督任务，在留言数据量较大，内容较为复杂的情况下，留言之间往往缺少区分度。例如表 5 所示，在第二步初步聚合而成的 701 类中，有 5 条留言被聚集成了一类：

留言编号	留言主题	留言详情
289834	A7 县文体中心乒羽中心何时能建成开放?	请问 A7 县文体中心乒羽中心何时能正式对外开放……
281722	建议 A7 县文体中心乒乓球馆尽快对外开放	A7 县东六路的文体中心乒乓球馆消防整改几个月了，……
192043	A7 县文体中心乒羽中心何时能建成开放?	请问 A7 县文体中心建成好几年了，一直听闻建有乒羽中心，……
258902 异常成员	请问非电力局的人能否在 A 市电力局的体育馆内打球?	您好！C 市电力局建了一高大上的室内体育馆，……
234468 异常成员	A 市贺隆体育馆是否有启用的计划安排?	贺隆体育馆总是闲置在那里也不是个事吧……

表 5　AP 聚类算法初步聚类后的某一类示例

可以看出，编号为 289834、281722、192043 的留言反映的是同一类问题，都是关于 A7 县文体中心的开放问题。而编号为 258902、234468 的留言，本来反映的问题与 A7 县文体中心无关，但由于它们也涉及了体育馆的内容，所以单纯按照语义向量的相似度，它们就容易与前 3 条留言混为一类。对于留言聚类于热点挖掘任务而言，它们实际上是异常值。

考虑到陈述一个热点问题，往往会涉及人名、地名、机构名等专有名词，我们提出一种基于 BERT 模型命名实体识别的留言聚类异常值过滤方法。命名实体识别（Named Entity Recognition,NER），就是指计算机从文本中自动识别出人名、地名、机构名等专有名词的过程。具体的程序实现，参考在 NLP 领域广泛运用的 python 工具包 HanLp[3]。工具包提供了基于 BERT 模型的 NER 实现接口。

我们利用 HanLp 工具包，识别了每条留言主题和详情中的命名实体后。设定过滤条件集合 R，在已完成 AP 初步聚类的留言中，凡是满足以下 3 个条件之一的留言，都将从所属类别中被筛除掉。

[3] https://github.com/hankcs/HanLP

三个过滤条件的集合 R：

条件（1）：留言的命名实体数量为 0。

条件（2）：留言所拥有的命名实体，在所属类别的全部命名实体集合中，都仅出现了 1 次。

条件（3）：留言所属类别最高频的 3 个命名实体，在该留言中均未出现。

在具体实现上，对于留言中的市县一级地名，如“A 市”、“A7 县”、“A3 区”，由于其出现频繁，降低了留言之间的区分度，我们不把它们纳入本任务的考虑之中，只在下文 3.5 节中生成热点问题表时使用。仍以表 5 的留言为例，表 6 显示了它们命名实体识别的结果，以及之后的过滤结果：

留言编号	留言主题	留言详情	识别出的命名实体（不含市县）	过滤结果
289834	A7 县文体中心乒羽中心何时能建成开放？	略	文体中心 乒羽中心	3 条件均不满足，保留
281722	建议 A7 县文体中心乒乓球馆尽快对外开放	略	文体中心	3 条件均不满足，保留
192043	A7 县文体中心乒羽中心何时能建成开放？	略	文体中心 乒羽中心	3 条件均不满足，保留
258902	请问非电力局的人能否在 A 市电力局的体育馆内打球？	略	电力局	满足条件（2），筛除
234468	A 市贺隆体育馆是否有启用的计划安排？	略	贺隆体育馆	满足条件（2），筛除

表 6 基于留言命名实体的异常值过滤示例

如表 6 的示例，根据命名实体的过滤条件集合 R，聚类数据得到了清洗，许多聚类的异常值被筛除，类别中成员的一致程度大大提高。

3.4 热度评价指标 L 设计

在筛除了异常值之后，我们设计了一套热点问题评价指标 L，针对清洗后数据中的每一个类别，分别计算其热度指标。热度评价指标主要考虑了如下 3 类因素：

（1）某聚类类别下的留言数量 n。同问题下的留言数量，是该问题热度的重要表现。

（2）某聚类类别下，最早留言日期与最晚留言日期的间隔天数 m。热点问题往往在较短时间内集中产生。

（3）每条留言的点赞数 a 与反对数 b。留言的点赞或反对数越多，也反映出更多的关注度。

则某聚类类别的热度 L = $\frac{\sum_{i=1}^{n}(10+\frac{a+b}{5})}{1 + \log_2(m+5)}$

3.5 实验结果分析

我们用 python 语言实现了 3.4 节所述的问题热度评价指标。得到了热度前 5 位的热点问题（详表请参见提交的文件“表 1-热点问题表.xls”），如表 7 所示。其中的地点/人群字段，由 3.3 节所述的 BERT 模型识别各条留言中的命名实体所得。问题描述字段，则通过抽取该热点问题下“点赞+反对数”最高的留言的主题内容，稍加润色所得。

热度排名	热度指数	地点/人群	问题描述
1	70.37	A4 区 p2p 公司	A4 区 58 车贷诈骗案
2	59.02	A 市 A5 区五矿万境 K9 县	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
3	42.33	A4 区 绿地海外滩小区 长赣高铁 渝长厦高铁	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？
4	31.29	A6 区 月亮岛路	关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉
5	23.18	A 市 富绿物业 丽发新城 暮云街道	A 市富绿物业丽发新城强行断业主家水

表 7 热度前 5 位的热点问题表

以热度排名第 4 位的“月亮岛路沿线架设 110kv 高压线杆”的问题为例，经上述的 AP 聚类与过滤异常值的步骤，该问题下的留言明细如表 8 所示。限于篇幅，表 8 仅展示每条留言的编号与主题。详表请参见提交的文件“表 2-热点问题留言明细表.xls”。

留言编号	留言主题
262052	关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉
272089	关于 A6 区月亮岛路 110kv 高压线的建议
218442	A6 区月亮岛路架设高压电线环评造假，谁为民众做主
225849	反映 A 市金星北片 110kv 及以上高压线的现状和规划的几个问题
254865	关于 A6 区月亮岛路沿线架设 110kv 高压电线杆的投诉
200480	关于 A 市金星北片 110kv 及以上高压线的现状和规划的几个问题
268250	关于 A6 区月亮岛路沿线架设 110KV 高压电线杆的投诉
234885	A6 区月亮岛路 11 万伏高压线没用地埋方式铺设
231773	反对 A6 区月亮岛路架设高压电线，强烈要求重启环境评估

表 8 热度位居第 4 的热点问题留言明细的示例[4]

由表 8 的示例可以看出，对于“月亮岛路沿线架设 110kv 高压电线”的热点问题，留言聚类的效果良好，9 条留言中没有异常值，具有较高的一致性。

4 多特征融合的答复意见质量评价

4.1 任务介绍与实验数据集

针对群众在网络问政平台上发布的留言，相关部门对留言反映的问题予以处理，并对留言进行答复。自动地、全面地评价答复意见的质量，有利于督促政府工作人员提高办事效率与水平，提高人民群众的满意程度。附件 4 共包含 2816 条记录，每条记录包含留言编号、留言用户、留言主题、留言时间、留言详情、答复意见、答复时间共计 7 个字段。我们拟选取多方面的特征，设计一套答复意见自动评价方案。

[4] 详表请参见提交的文件“表 2-热点问题留言明细表.xls”。

4.2 答复意见的质量评价指标

答复意见的质量评价指标，将综合参考答复与留言的相关性、答复的完整性、答复的可操作性、答复的条理性等维度。为此，我们提出了如下 9 项特征。除了答复的时效程度为扣分机制，其余 8 项各自的满分均为 100 分，再按照不同的权重进行加权求和。

（1）留言中的命名实体，在答复中的覆盖率（权重 20%）。满分 100 分。如第三章所述，人名、地名、机构名等命名实体，是留言所反映问题的重要特征。答复意见对这些命名实体的覆盖率，也反映了答复与留言的相关程度。

（2）答复与留言的语义向量相似度（权重 20%）。满分 100 分。答复与留言都针对的是同一个问题，因此或多或少在语义上会有一定的相似度。该指标可以防止“答非所问”、“离题万里”的情况发生。

（3）答复的长度（权重 20%）。答复意见的长度，能一定程度上反映答复的细致程度与全面性。长度过短的答复必然质量不高。例如附件 4 数据集中，编号为 6556 的留言询问了“打狂犬疫苗报销比例是多少”，这本是一个事实性的问题，然而答复却仅仅有“已收悉”3 个字。我们规定，答复的长度若低于 30 字，计 0 分；超出的字数按 $30 + \frac{超出的字数}{10}$ 计分。满分为 100 分。

（4）答复中关键词的覆盖率（权重 20%）。满分 100 分。基于对数据的观察，我们发现，质量较高的答复通常含有较多的“关键词”。这些关键词中，有的反映出政府部门对问题高度重视的态度；有的则反映出政府部门采取的积极措施行动。我们归纳总结的 91 个关键词如表 9 所示。

答复	依法	咨询	收悉	调整	保证	及时	反映	办理	处理	解决
通知	开展	电话	拨打	详询	询	应该	支持	商议	改造	监督
理解	督促	规定	工作	建议	意见	尽快	核实	建设	鉴定	调查
查	研究	积极	加强	力度	确保	要求	论证	确认	贯彻	落实
查处	整改	办理	核查	执法	整治	检查	指导	根据	提供	务必
跟进	检测	按照	核定	行动	查处	严格	把关	保障	巡查	重视
规划	调查	查明	部门	行政	协议	处置	妥善	请求	报告	争取
按照	请示	获批	受理	统筹	会商	建设	解释	劝诫	责令	热线
联系	取缔	打击								

表 9 答复的关键词表

（5）**答复中相关法律法规、政策文件的出现与否（权重 5%）**。针对群众反映的医疗、交通、劳动与社会保障、商贸、环境、卫生等问题，提供相关的政策文件、法律法规，有助于提升答复的专业程度与可信度。这可以由答复中有无书名号“《》”来判断，若有，记为满分 100 分；若无，记为 0 分。

（6）**答复中联系方式的出现与否（权重 5%）**。答复中若含有相关部门的电话号码联系方式，就为群众进一步咨询与处理问题提供了渠道。这也反映出工作人员广开言路的良好作风。答复中若出现电话号码，记为满分 100 分；若无，记为 0 分。

（7）**答复中相关网址链接的出现与否（权重 5%）**。这为留言群众提供了相关的参考资料，或是进一步解决问题的手段。答复的文本中若出现 http 或 https 的字符串，记为满分 100 分；若无，记为 0 分。

（8）**回答的条理性（权重 5%）**。若答复文本中出现了诸如“一、……二、……三、……”、“（一）、（二）、（三）”、“（1）……（2）……（3）……”、“1.……2.……3.……”的序列标志，或包含“首先”、“下一步”、“最后”等关键词，记为 100 分，否则记为 0 分。

上述 8 项指标的权重之和为 100%，再考虑第 9 项指标：答复的及时性。

（9）**答复的间隔时长**。这反映出答复的及时程度。留言发布的 15 天内答复，不扣分，超出的每 1 天扣 0.2 分。

4.3 实验结果分析

我们使用 python 语言，实现了上述“9 项合一”的评价指标，分别对 2816 条留言质量进行了计算。如图 7 所示[5]：

[5] 详表请参看提交的文件“06 答复意见质量评价结果.xls”。

A1 答复对应的留言编号

	A	B	C	D	E	F	G	H	I	J	K
1	答复对应的!	最终质量得分	答复与留言的相似度	答复的长度分	命名实体的覆盖率	关键词覆盖率	法律法规分	条理分	网址链接分	联系方式分	扣除的时效分
2	24793	81.62178917	96.4453095	87.8	100	23.86363636	100	100	100	100	0
3	119178	79.1448523	95.72426148	100	100	25	100	100	100	0	0
4	4331	78.99301253	97.23778992	100	100	22.72727273	100	100	0	100	0
5	7482	78.59902728	95.26786366	100	100	22.72727273	100	100	0	100	0
6	88359	77.09167184	96.82199558	100	100	38.63636364	100	100	0	0	0
7	96757	77.05663475	96.64681013	100	100	38.63636364	100	100	0	0	0
8	16719	75.98135714	96.95224024	100	100	32.95454545	100	100	0	0	0
9	32722	75.94265559	92.21327795	100	100	12.5	100	100	0	100	0
10	25655	75.4544829	93.37241452	96.4	100	12.5	100	100	0	100	0
11	115234	75.44782404	96.10275657	100	80	26.13636364	100	100	0	100	0
12	133336	75.42313355	94.16112232	100	100	32.95454545	100	100	0	0	0
13	99213	74.72666005	94.0878457	100	75	54.54545455	100	100	0	0	0
14	23009	74.68432761	95.01254715	100	100	28.40909091	100	0	0	100	0
15	110671	74.65875144	96.02102995	100	100	27.27272727	100	100	0	0	0

图 7　答复质量评价结果示例

以获得最高得分 81.6217 的答复留言为例，其答复的全文如下：

网友：您好!您于 2019 年 1 月 22 日咨询的关于“投诉 A7 县宁华置业星湖湾(洋房）二期未达标且强行交房”的问题，我局已收悉，经核实，现回复如下：一、关于您提出的楼间距设计不达标的问题：根据已批复的 K8 县·星湖湾二期总平面图，该小区 1#与 7#、8#（https://baidu.com/xxxx）楼间距为 19.5m，满足《A 市城市规划管理技术规定》2009 版 ii 类地区居住建筑平行布置 1.1h 建筑间距要求；二、关于您提出的走廊过道施工不达标的问题:已要求建设单位联系原设计院和施工单位共同到现场查看，出具书面报告情况，根据书面报告再做进一步处理；三、关于您提出的 3 号栋与电梯相邻卧室、起居室未做隔音处理问题:已责令建设单位进行整改；四、关于您提出的开发商各栋 1 楼大厅入户门截止 2019 年 1 月 15 日未完工，消费水管未通水，绿化未完成，涉嫌走不正当程序拿到了完成竣工验收，甚至拿到了不动产证的问题：该项目于 2018 年 12 月 21 日由五方责任主体对 1、2、3、8、9 栋及商业 s01、s02 栋进行了竣工验收。根据《建设工程质量管理条例》第十六条规定：“建设单位收到建设工程竣工报告后，应当组织设计、施工、工程监理等有关单位进行竣工验收，并签署质量合格文件后，视为验收合格。”其绿化景观等室外附属设施也已完成，能够保证户主的正常入住。如有其它问题需要咨询，建议拨打 0000-00000000。感谢您对我们工作的理解和支持！2019 年 1 月 29 日

图 8 质量分数最高的答复意见

如图 8 所示，该答复意见内容详实；答复共分为四点，条理清楚；说理时引用了相关的条理、规定，说服力和专业性强；提供了相关网址链接作为参考资料；最后还留下了联系热线，供用户进一步的咨询。经人工评价，它确实属于质量较高的答复意见。

5.结语

本文基于近年来语言智能的里程碑技术：BERT深度语言模型，围绕“智慧

政务”文本挖掘的主题，完成了留言自动分类、热点问题挖掘和答复意见质量评价3项工作。

在留言自动分类任务上，BERT模型的F-score达到了92.89%，高出基线LSTM模型约11.6%，也高出Fasttext模型约1.8%。这证明了BERT模型在政务留言文本分类任务上优越性与实用性。

在留言热点问题挖掘任务上，我们创新提出了首先基于Affinity Propagation的算法对留言进行初步聚类，再通过命名实体识别进一步过滤类别异常值的策略。这不需要事先预设聚类数目，且类中成员的一致性得到了显著提高。我们结合了话题时长、留言数、留言的点赞与反对数共3类因素，实现了综合评价每个问题（话题）的热度值。

在答复意见评价任务上，我们设计了9项特征综合评价答复意见：答复与留言的相似度、答复长度、命名实体的覆盖率、关键词覆盖率、法律法规分、条理分、网址链接分、联系方式分、扣除的时效分。多特征融合的评价体系有助于客观、全面反映答复意见的质量，促进政府工作人员改进工作态度，增强为人民群众排忧解难的能力。

在未来的工作中，（1）在留言自动分类任务上，我们将分析留言分类中的分类有误的例子，分析模型可能失误的原因。重点观察F1值在各个模型上都普遍不高的商贸旅游类留言。以总结规律，进一步改进留言分类模型。（2）在留言热点问题挖掘任务上，我们将尝试调整AP聚类算法的preference超参数，以尝试发现更优的聚类效果。（3）在答复意见评价任务上，我们考虑引入NLP的自动句法分析（parsing）或语义分析技术，衡量答复意见的语句通顺与语义连贯（coherence）程度。

参考文献

[1] Mikolov T, Sutskever I, Kai C, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing

Systems, 2013,26:3111-3119.

[2] Peters M E, Neumann M, Iyyer M, et al. {Deep contextualized word representations}[J]. arXiv e-prints, 2018:1802-5365.

[3] Devlin J, Chang M, Lee K, et al. {BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding}[J]. arXiv e-prints, 2018:1810-4805.

[4] Google. TensorFlow code and pre-trained models for BERT[EB/OL]. https://github.com/google-research/bert.

[5] Vaswani A, Shazeer N, Parmar N, et al. {Attention Is All You Need}[J]. arXiv e-prints, 2017:1706-3762.

[6] Sun C, Qiu X, Xu Y, et al. {How to Fine-Tune BERT for Text Classification?}[J]. arXiv e-prints, 2019:1905-5583.

[7] Hochreiter S et al. Long Short-term Memory[J]. Neural Computation, 1997(9(8)):1735-1780.

[8] Joulin A et al. Bag of Tricks for Efficient Text Classification: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, 2017[C].

[9] Zaremba W, Sutskever I, Vinyals O. {Recurrent Neural Network Regularization}[J]. arXiv e-prints, 2014:1409-2329.

[10]邓三鸿, 傅余洋子, 王昊. 基于LSTM模型的中文图书多标签分类研究[J]. 数据分析与知识发现, 2017,1(07):52-60.

[11]Brendan J. Frey D D. Clustering by Passing Messages Between Data Points[J]. Science, 2007,315(5814):972-976.

[12]刘晓勇, 付辉. 一种快速AP聚类算法[J]. 山东大学学报(工学版), 2011,41(04):20-23.