KGCN4Comp：中文投诉短文本处理机制研究

一、摘要

来自公众的大量投诉短文本如何进行高效的分类？这是政府部门应用智能算法后进行数字化政务和推动民意高效快速落实的关键。不同于一般的短文本分类任务，特征稀疏、语境复杂、表意能力差、主题多样、数据噪声高的投诉文本数据如何准确表示，在之前的研究中仍旧是一个难点。应对这种挑战所构建的 KGCN4Comp 包含一个拥有四层结构本体模式的知识图谱，用以表达整个中文投诉场景，和一个有自适应能力的图卷积神经网络，进行特征工程和分类。在基于输入文本生成的子图和投诉短文上，比较了 10 种特征计算方式和分类器组合在一个收集到的中文数据集上的分类性能。在中文文本处理领域之前的研究中，很少有研究者通过构建知识图谱和应用图神经网络的方法完成投诉文本的多标签分类工作。KGCN4Comp 在投诉短文本分类任务上体现出了明显的优越性，相较于 Bert 模型的分类准确率提升了 4.1%。

二、介绍

政府部门每天会接收到大量的投诉文本[1]，而这些来自公众的投诉文本如何进行高效、准确的分类是数字化政务推进的关键技术难关[2]。在传统的分类场景中，投诉文本多由政府部门的接线员通过人力进行标注和处理，极易因主观因素出现误判。于是近年来，基于特征工程和分类器的机器学习分类方法在政府部门投诉处理场景下的短文本分类任务上展现出了较好的效果[3]，能够达到一部分人工标注的效果。投诉文本有着特征稀疏、语境复杂、表意能力差、主题多样、数据噪声高等特点，在文本的特征计算上很难做到像长文本分类一样大量、有效地提取特征[4]。一种常见的方法是使用独热编码来对特征进行表达，以抵抗数据特征维度稀疏的现象[5]。但应用这种特征表示方法的分类精度依旧差强人意。在传统的文本分类任务中，特征计算往往使用基于序列或词袋的方法，但这在数据量小和特征稀疏的短文本分类中很难有好的表现。

在一个实验中（如图 1），我们观察了文本长度对于分类器性能的影响。事实上，我们使用 GPT4 对一个有 100 条数据的短文本数据集进行了随机扩增，使之成为五个对应具有相同表意但不同体量大小的数据集。在将语句看成序列和词袋的传统分类方式中，随着单个文本规模的变大，分类精度产生了明显的上升。同时我们还发现，在这样的尝试中传统方法的知识抽象在短文本数据集上的表现并不好。而投诉短文本的字数体量普遍在 100 字左右。

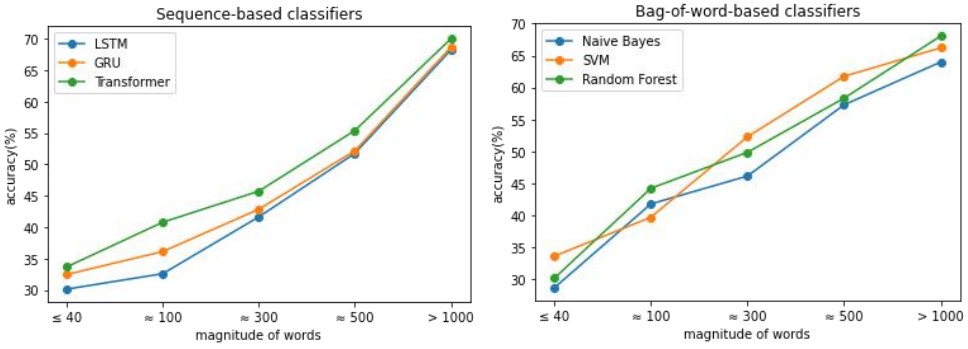

图 1 基于词袋和序列的分类算法对单个文本体量不同数据集的分类性能

我们提出的 KGCN4Comp 使用特种结构的知识图谱本体模式对现有知识进行表达，以增强整体语境的关键词感知和检索能力。应用知识图谱的知识表达方法在显式上精简了来自复杂语境的文字描述，计算时可以避免大量情感词汇和低价值词汇带来的特征偏差，解构和呈现了整段投诉文本最基本的语句构成。经大量工作验证，KGCN4Comp 在文本分类任务上是一种合理、高效的知识表达手段。在隐式上，知识图谱连通了不同语境的实体对象，增强了模型在处理陌生文本时的泛化能力和推理能力，并更进一步强调了节点级别上的可扩展性。我们为整个知识图谱设计了一个四层结构的本体模式，以增强整个图对知识的表达能力和可解释性。借助知识图谱对多源异构数据的强表达性，我们整合了大量相关数据以补充单一投诉文本对整个投诉处理工作流程描述上的欠缺，以“问题引发现象”、“现象导致结果”、“结果指向投诉类别”、“投诉类别决定处理部门和方案”的思路，构建了逻辑清晰的层内和层间关系链路。

KGCN4Comp 使用的节点特征计算方法是基于相邻节点的，具有天生的扩展属性和知识关联能力[6]。理论上，只要知识图谱足够丰富、体量足够大，基于此的文本分类性能将会无限接近最优。本文在子图的特征计算中所采用的跳步计算机制也可以在一定接受范围内扩大相邻节点的感受野，以增强子图的语义表达。

另外，由于研究场景的多变性，现有的很多研究并不是基于相同的样本空间和基准数据集的，这将导致讨论的不完全和难比较性。我们使用政府部门提供的投诉工单分类方法，构建了适用于整个中文投诉场景的类别体系，并基于此提出了一个可供参考的中文投诉短文本分类数据集（GCD），这个数据集可以在以下链接找到①。

为了完成这一任务，并以此提高政府部门对投诉数据的分类效率，我们基于知识图谱和自适应特征嵌入的图卷积原理构建了一个称为 KGCN4Comp 的算法框架，用于子图的特征计算和分类。分类结果可以作为新的输入在知识图谱中获取关键字和相关节点。将其输入一个基于 GPT3.5 的问答平台进行投诉的处理和反馈。据我们所知，本研究的贡献如下：

（1）针对行业中欠缺专业数据集的现状，根据政府提供的分类模式，整理和形成了高质量的数据和标签，并提出了首个中文投诉短文本分类数据集（GCD），为后来的投诉文本分类研究建立了良好的实验平台。

（2）针对复杂语境和单一数据来源对整个投诉文本处理工作的表意不全，提出了一种四层的本体模式。该模式合理地将异源数据表达为多层，有效沟通了投诉文本和政务处理流程之间的关系，为应用知识图谱的投诉短文本多标签分类领域研究在中文语境下进行了首个实验性尝试，并提供了一个本体模式构建的典型范式。

（3）为了降低民众的使用门槛和操作难度，我们基于 GPT3.5 构建了一个用户友好的问答系统，该系统将输入的文本数据自动转化为子图，在预训练的知识图谱上计算特征并输出关键字，经由大模型生成沟通体验良好的对话，满足了公众对投诉问答系统的易用性、友好性需求。

这篇文章的剩余部分如下。第三节介绍了中文短文本多标签分类领域的有关工作，第四节介绍了 KGCN4Comp 的理论基础，第五节我们描述了一组基于 KGCN4Comp 的实验，第六节总结了本研究并提出了对未来发展的展望，第七节是引用文献。

三、相关工作

短文本分类深度学习方法 深度学习方法以其优秀的分类性能易训练性在以往的分类任务中大放异彩[7,8]。其在对长文本的理解上有令人印象深刻的成果[8,9]（深度学习的一些先进成果），一些领先的研究应用深度学习模型对上下文的感知能力[10,11]，实现了好的模型精度。但是对于短文本的理解和特征工程，通常体现出较低的评分，这通常来自短文本的特征稀疏性[12]。一种典型的办法是对文本进行扩展[13]，但这面临着苛刻的文本理解任务，否则极易引入歧义。本文中使用的知识图谱模型在这方面实现了实体消歧和实体扩增的目的，可以很好的应对这种问题。一种基于 KG 和 GCN 的方法自提出就受到了研究者极大的关注和应用[14]，使用 KGCN 的推荐系统在稀疏性和冷启动问题上获得了很大的改善。基于这个优秀机制的启发，KGCN4Comp 能够在投诉文本的分类上获得了更好的性能。

基于知识图谱的文本知识表达 知识图谱对多模态异源数据友好，能够进行广泛的知识扩展和推理，使得它在文本理解中有很强的适应性。其知识表达能力在可扩展性和可解释性上表现出众[15]。使用图结构对投诉短文本进行表达拥有天然的优势，因为包含情感词汇和低表意能力语句的投诉文本可以借助一个高精度的实体关系抽取模型获得主题增强和数据过滤[16]。

多标签文本分类 涉及复杂应用场景的实体可能具有多个属性，这对分类器的分类性能提出了更高的要求[17,18,19]。单标签分类关注在样本空间中对象的指向精度[20]，这是由特征表示、模型选择和参数调整、训练数据量以及优化目标和损失函数等因素共同决定的。而多标签分类任务不仅要考虑上述问题，还要考虑标签之间的映射关系[21,22]。

四、KGCN4Comp 原理

4.1 面向投诉短文本的知识图谱构建

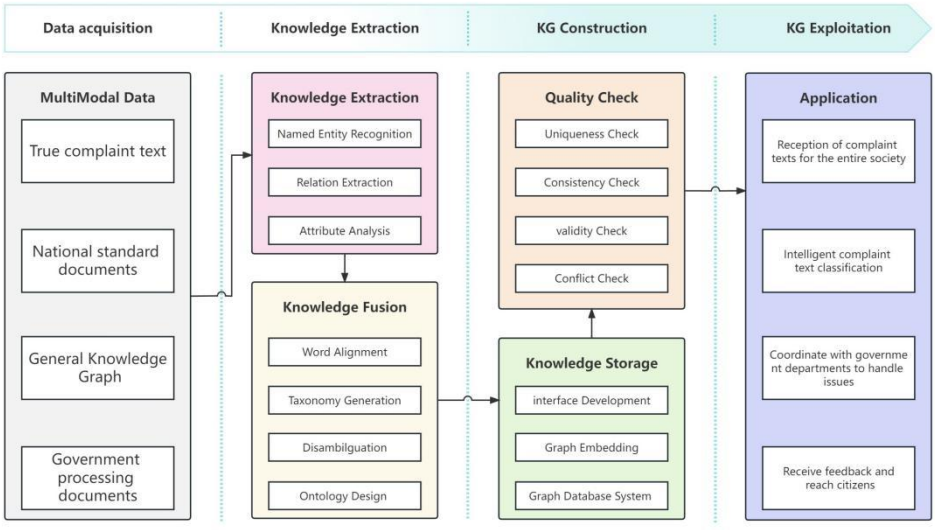

图 2 面向投诉的知识图谱构建流程

多源数据的抽取、整理和组织是海量数据能够用于投诉短文本分类任务的关键[23]。面向投诉文本的知识图谱描述了被投诉的现象、产生的影响、所属的分类和处理方法。因此应当包含但不限于表 1 中标识的节点类型及图 4 中边的情况。

表 1 知识图谱节点设计

节点名称	描述
事务节点	投诉工单中的实体
问题节点	实体引发的问题
引发后果节点	问题带来的影响和后果
处理流程节点	有关部门对问题的处理流程
一级类别	分类树上的一级节点
二级类别	分类树上的二级节点
三级类别	分类树上的三级节点

面向投诉文本的知识图谱所维护的知识通常是异构的，包括大量原始投诉文本、政府工作的处理意见、国民经济行业分类注释[24]、有关规章制度和类别样本空间等数据。原始数据中存在的非结构化数据，比如投诉工单文本，政府部门对投诉的处理规范等，被整理和抽取成三元组的形式，参与最终知识图谱的构成。文本类别空间和国民经济行业分类数据则是以一种树形结构进行存储，经过转化也形成三元组参与整个知识图谱的表达。应用图 2 中设计的流程，经过处理的三元组数据可以构建成为图 4 表达的形式。

在我们提出的知识图谱中，三元组的“实体-关系-实体”模型是根据实际情况生成和标注的，具有精准的表意。另外，在可接受约束下，一定程度的信息冗余对整个模型来说是有利的，这能提供更丰富的知识基础[25]，但其中不正确的连接和因为同形异义词而造成的错误，应当使用实体对齐和实体消歧的手段去除[26]。

另外，对于重要的实体数据，例如“企业实体名称”和“个体实体名称”（通常是一家公司或个人商铺的全名，例如“ 山东汇宠生物有限公司”），我们设计了一个精准的行业类型抽取模型，可以判断其所属的行业类别，然后根据“国民经济行业分类注释”对其实体关系进行了增广，将这类数据映射到更旷阔的实体空间上，如图 3 所示。具体方法是，我们构建了一个由 700 个企业行业词汇组成的词库，为关键字识别算法提供语料，使用了一个分词组件和匹配机制，为企业/个体实体名称进行了行业类别的标注。在临淄市的 45300 个个体实体名称和 23037 个企业实体名称数据集上得出的实验结果表明，这套方法的精度在 96%以上。

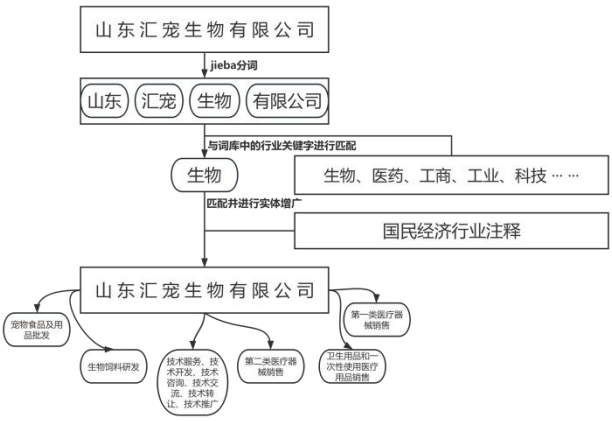

图 3 企业/个体实体增广

为了更详细的说明这个规则范式，我们使用了 GCD 数据集中的一条真实数据，在图 4 中翻译成英文后进行了示例。面向投诉的知识图谱可以分为四层。事务层存储着由投诉文本中抽取出来的三元组所组成的子图，它详细描述了在一般情况下投诉人对现实

情况的反馈，具有复杂且丰富的知识表意。同时我们对其中的重要实体节点进行了扩增，使得其表达更加准确。描述层包含了现实情况（事务层）引发的问题和结果，这加深了现象到投诉分类的归因。类层维护了一个政府投诉处理部门提供的完整投诉工单分类标准，如图 3 中所示，诉求类别和客体分类分别是两个树形结构，包含了投诉文本的所有类别属性。执行层描述的是政府部门将不同类型工单配送给不同管理部门进行处理的过程，其中的数据抽取自各级下游投诉处理部门或管理部门的工作章程文件和指导文件。

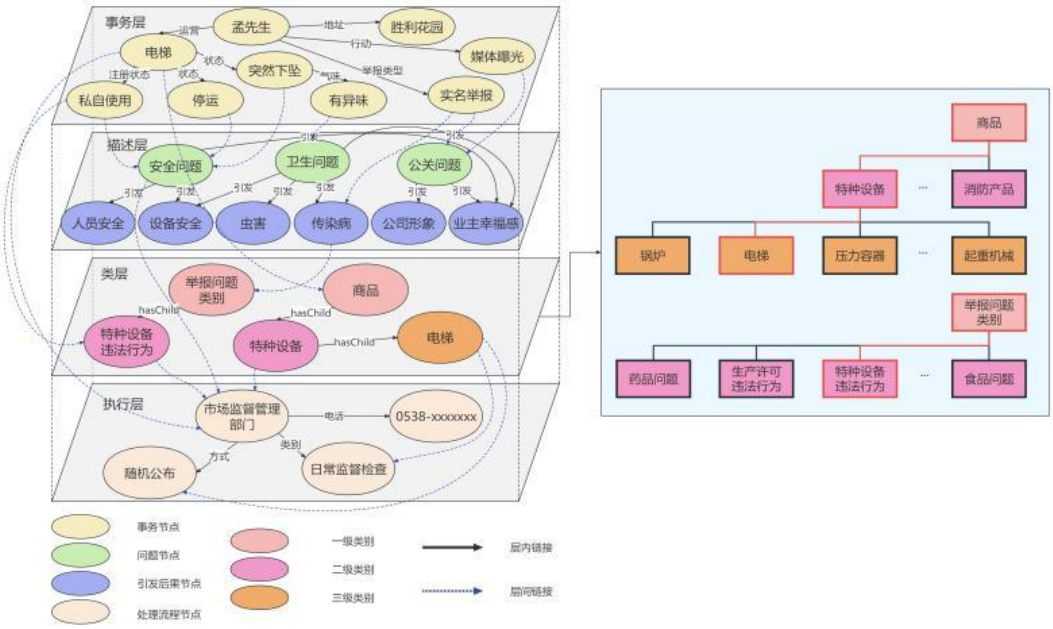

图 4 面向投诉的知识图谱结构示意

4.2 自适应的图特征工程

4.2.1 邻接矩阵和节点特征初始化

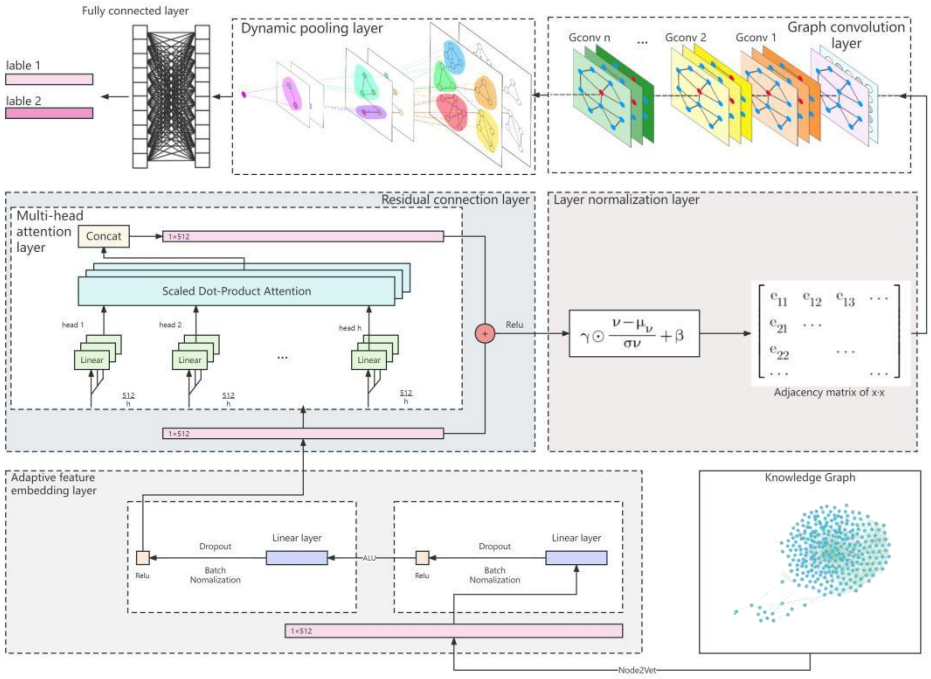

图 5 图神经网络算法逻辑图

图神经网络的基本原理是将图的连通状态视为先验知识，以此来计算整个图中每个节点的特征[27]。而在短文本分类任务中，我们计划使用一个具有自适应层的图神经网络来计算整个知识图谱中节点的特征向量[28]，以此来描述某个实体在整个投诉语境中的重要度。

在自适应特征嵌入层中，我们设计了一个应用 node2vec 算法的模块完成特征嵌入[29]，计算了各个节点大小为 1×512 维的初始特征向量，并希望在后续的算法过程中处理和优化这些特征。设计有两组相同的算子对特征向量进行处理和计算。其中线性层用于全局特征处理，将输入的一维向量进行初步转换；Dropout 通过随机丢弃一些神经元的输出，有效地避免过拟合问题，提高模型的泛化能力；Batch Normalization 通过对每个 mini-batch 的输入数据进行归一化处理，可以有效地提高训练的稳定性和收敛速度，同时也可以起到正则化的作用。

为了对自适应特征嵌入层输出的特征进行进一步的处理，以更好地捕捉特征间的关联性和重要性，我们设计了一个多头注意力机制层[30]。

多头注意力机制可以将输入特征进行分组，每组都通过自注意力机制计算出一个权重向量，然后将所有组的权重向量进行加权平均得到输出的特征向量。通过这种方式，多头注意力机制可以更好地提取输入特征中的相关信息。

$$MultiHeadAttention(Q,K,V) \quad = Concat(head_1, head_2, \dots, head_h) \cdot W_O \; head_i$$

其中，Q、K 和 V 分别表示查询、键和值的输入向量序列，h表示注意力头的数量，W_Q^i、W_K^i和W_V^i是对应于第i个注意力头的线性变换权重矩阵，W_O是拼接后的向量的线性变换权重矩阵。

残差连接层是一种用于解决深度神经网络训练过程中梯度消失问题的技术。在残差连接层中，模型学习的是残差（即目标输出与当前输出之间的差异），而不是直接学习目标输出。这使得训练深度神经网络更加容易，因为它可以减少梯度消失问题。在多头注意力机制层后加入残差连接层的作用是使得模型可以更好地捕捉输入的特征。因为残差连接层允许模型跳过一些层，从而减少信息的丢失，使得模型可以更好地保留输入的特征，从而提高模型的准确性和可解释性[31]。

$$ResidualLayer(X) \quad = X + MultiHeadAttention(X,X,X)$$

其中，X表示输入特征向量，$MultiHeadAttention(Q,K,V)$表示多头注意力机制的输出，该输出使用输入特征向量$X$作为查询、键和值。

在层规范化层中，我们在不同的数据尺度上进行标准化[32]，通过预设的标准差和均值完成整个特征向量的归一化操作。

$$LayerNormalization(X) \quad = \gamma \odot (\frac{X-\mu}{\sigma}) + \beta$$

其中，X表示输入特征向量，μ和σ分别表示输入特征向量的均值和标准差，γ和 β是可学习的缩放参数和偏移参数。

$$DiagonalTensor(X) \quad = diag(X)$$

其中，X表示层规范化层的输出结果（一个 512 维的向量），$diag(X)$表示将向量 X的元素按顺序放在一个对角矩阵的对角线上，形成一个张量。

至此我们可以得到一个存储着整个知识图谱连接情况的邻接矩阵，矩阵的正对角线上存放着节点的特征向量，后续的操作就是在训练过程中针对这个矩阵进行维护和更新，使得新计算出的矩阵能够支撑更好的分类精度。

4.3 子图的生成与增广

对于每一个被输入的投诉文本，算法都要将其抽取和转化成为描述一个中心语义的连通子图，如图 6。于是，计算一个文本的特征向量就被转化为了计算一个子图的特征向量，这对于图结构来说是容易操作的。

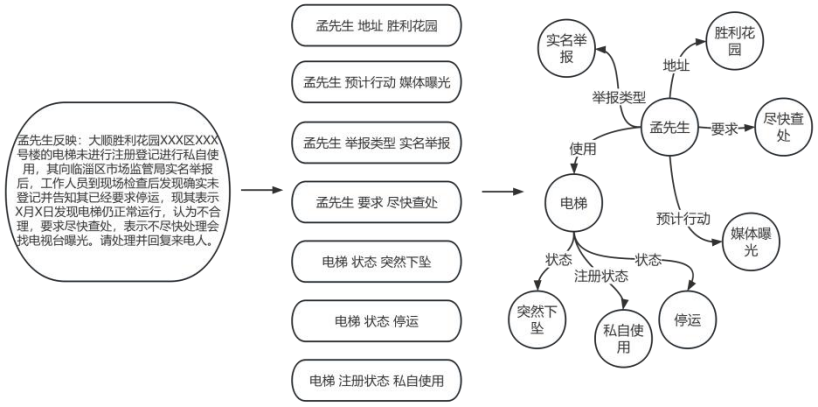

图 6 投诉文本的实体抽取与子图构建

在实际训练中我们发现，一个稀疏子图的特征计算需要大量相关节点的参与以克服特征稀疏的向量表示。一种常用的方法是针对图中连通的其他节点进行跳步延长和实体增广，使得更多的节点被纳入到子图特征的计算范围中来。在我们设计的图神经网络中，子图扩增共分为两步，1）根据子图生成的节点，查找知识图谱中的相同或相关节点。2）根据有关节点对子图进行扩增。

在 1）中我们使用了一个校验文字相关性的模块，用于构建输入节点和知识图谱中节点之间的关系。具体模式为：根据实体关系抽取模型，构建数据文本的子图模式，搜索预训练的所有节点中该对应节点的特征向量；对于知识图谱中未出现的节点，以子图中该节点的邻接节点的特征为基准，计算这些节点各维度的均值作此类节点的特征向量。

在 2）中，我们回溯整个知识图谱，使用一个半随机的跳步延长操作扩增整个子图的实体节点数，即在特征距离满足一定阈值的临界节点中选择延申方向随机扩展，这表达了一种期待语义上相邻的节点可以被找到并加入到整个子图中来的愿景。这样做的好处是能够扩大特征计算时以子图节点为核心在整个知识图谱中的感受野，但是过多杂乱的相关信息会干扰分类器的判断能力。所以我们在扩增过程中控制了新节点的随机选择，加入了一个根据信息增益和节点重要性的模块，用来判断新加入的节点是否在知识表达上为整个子图带来了帮助。当这个模块无法断定剩余节点是否应该被纳入子图时，重新执行随机选择。

信息增益（Information Gain，IG）： 我们可以将图的聚类系数作为熵来计算信息增益[33]。图的聚类系数 C 描述了图中节点的聚集程度，它被定义为：

$$C = 3 * Num_{triangle}/Num_{triples}$$

其中，$Num_{triangle}$是图中所有完全连接的三个节点（形成一个三角形）的数量，$Num_{triples}$是图中所有连接的三个节点（不一定完全连接）的数量。

假设当前的子图的聚类系数为 C_{before}，添加新节点 N 后的子图的聚类系数为 C_{after}，

则信息增益 $IG(N)$ 可以定义为C_{after} 和 C_{before} 的差值，即

$$IG(N) = C_{after} - C_{before}$$

节点重要性（Importance, Imp）：我们可以将节点重要性定义为接近中心性（Closeness Centrality）。接近中心性是一种描述节点在网络中的中心位置的度量，定义为所有节点到该节点的最短距离的倒数的平均值[34]。

对于图 $G=(V,E)$，节点 v 的接近中心性 $C(v)$ 计算公式如下：

$$C(v)=1/\sum_u(d(u,v))$$

其中，$d(u,v)$ 是节点 u 和 v 之间的最短路径长度，$\sum$ 是对图中所有其他节点 u 的求和。

因此，对于节点 N，其重要性就是其在图中的接近中心性，即

$$Imp(N)=C(N)$$

然后，我们可以通过一个加权函数来计算每个待添加节点的得分这个得分可以决定一个节点的所有邻接关系中，哪个节点应该被添加到子图中：

$$Score(N)=w1*IG(N)+w2*Imp(N)$$

另外，在扩增子图过程中，跳步机制的最大深度是需要考量的重要参数。如图 7，我们在一百条数据的小数据集中进行了子图增广操作，对于不同增广深度的数据进行了特征计算，以最终文本的分类正确率作为指标。在这个中文投诉场景的实际操作中，我们发现算法性能在层数在 7 处取最优。

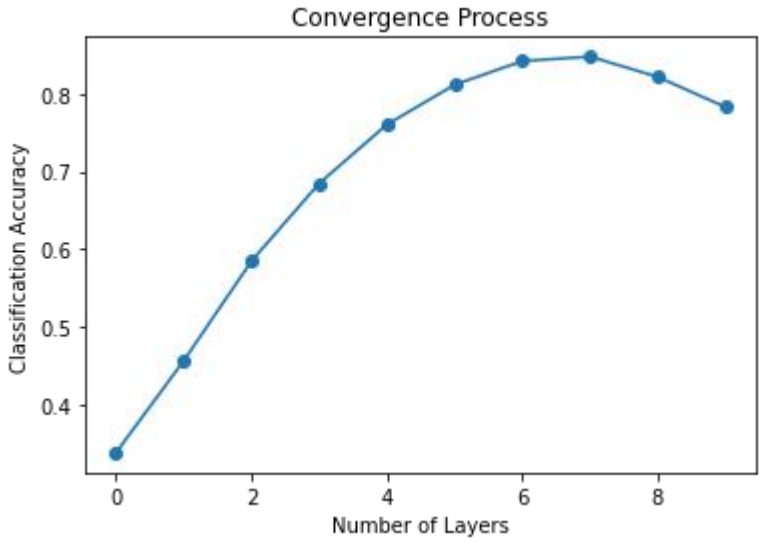

图 7 子图扩增的层数与分类准确性的关系

4.4 GCN 分类器训练

图卷积神经网络（Graph Convolutional Neural Network，GCN）是一种输入是图数据的深度学习模型[35,36]。它在图结构中利用节点和边的信息来学习节点的表示，并在此基础上进行图分类任务。在本研究中，图神经网络的输出是一个多标签的复合形式，对于一个投诉文本，分类器需要指出其诉求类型、诉求类别和客体分类三个属性。在后续的实验中，模型对于一个样本的任意一个标签分类错误的情况一律判负。

我们设计的 GCN 训练流程如下：

构建输入图：将文本分类问题转化为一个图结构。每个节点表示一个词语及其特征，边表示它们之间的关系。

图卷积层：利用图卷积神经网络来学习节点的表示。GCN 通过聚合节点的邻居信息来更新节点的表示。每个图卷积层的更新公式如下：

$$H^{(l+1)}=\sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

其中，$H^{(l+1)}$是第l层节点的表示矩阵，$\hat{A}$是归一化的邻接矩阵，$\hat{D}$是度矩阵的对角线上元素为度的平方根的矩阵，$W^{(l)}$是第l层的权重矩阵，σ表示激活函数。

动态图池化层：利用动态图池化层来对图中的子图进行池化，以提取关键信息。动态图池化通过自适应地选择重要节点，并对其进行聚合来形成池化子图[37]。

假设输入是一个图 G,其中$H^{(l)}$表示第l层的节点表示矩阵,$S^{(l)}$表示节点的分数矩阵，$P^{(l)}$表示节点的池化概率矩阵。

计算节点分数：

$$S^{(l)} = ReLU(H^{(l)}W_s^{(l)})$$

其中，$W_s^{(l)}$是分数计算的权重矩阵，$ReLU$是激活函数（如 ReLU）。

计算池化概率：

$$P^{(l)} = softmax(S^{(l)})$$

将分数矩阵经过 softmax 函数得到池化概率矩阵，使得每个节点的概率值都在 0 到 1 之间且总和为 1。

节点池化：根据池化概率矩阵$P^{(l)}$，选择部分节点进行池化。

分类层：将池化后的子图输入到一个全连接神经网络中，最终输出分类结果。

五、实验

5.1 数据集

为了推动中文投诉短文本分类的研究，我们根据收集到的真实投诉数据，形成了基于场景多样性的 GCD（Government Complaints Data）数据集。这个数据集有以下特性。

场景多样性 GCD 数据集的主要贡献是提出了自然场景下较为完整的中文投诉文本分类体系，能够囊括所有可能出现的投诉类型。体现在数据集中，GCD 数据集在包括但不限于“食品安全投诉”“质量投诉”“合同投诉”“售后服务投诉”等在内的多个大类，精准地标注了投诉文本的诉求分类和客体分类标签。

表 2 GCD 数据集场景分布

数据集	数据体量	描述场景
GCD-FSC 政府部门投诉工单数据-食品安全投诉	1097	对食品质量或安全问题提出的不满和意见的投诉
GCD-QC 政府部门投诉工单数据-质量投诉	929	对产品质量存在问题的不满和意见的投诉
GCD-CC 政府部门投诉工单数据-合同投诉	454	因合同履行中的问题或纠纷而提出的不满和意见的投诉
GCD-ASC 政府部门投诉工单数据-售后服务投诉	534	对购买产品后所遇到的服务质量问题提出的不满和意见的投诉
GCD-UCC 政府部门投诉工单数据-不正当竞争投诉	179	对违反公平竞争原则的商业行为提出的不满和意见的投诉
其他类别	1119	其他场景下的投诉数据

多模态数据 基于知识图谱的文本表示一直使用“三元组”的方式抽象语句中的主谓宾。而在之前的研究中，被广泛使用的“实体，关系，实体”和“实体，属性，属性值”范式是实体关系抽取任务的主要研究对象。在原始文本以外，数据集中的每个主题我们都提供了一个以三元组格式组织的数据集与 GCD 数据集中的文本相对应，以便使用者对实验模型的抽取能力和精度进行评估。这些三元组的生成，一小部分来自 chatGPT 的理解，更多的工作量来自于人工校正和基于联合解码的关系抽取模型。

在之前的研究中我们并没有发现高质量的、面向政府部门的投诉文本数据，所以在模型验证中，我们在收集到的 GCD 数据集上进行了广泛的实验。

实验中主要使用了 GCD 数据集中的四个主要组成部分，分别是 GCD-FSC（政府部门投诉工单数据-食品安全投诉 Government Complaints Work Order Data - Food Safety Complaints）。GCD-QC（政府部门投诉工单数据-质量投诉 Government Complaints Work Order Data - Quality Complaints）。GCD-CC（政府部门投诉工单数据-合同投诉 Government Complaints Work Order Data - Contract Complaints）。GCD-ASC（政府部门投诉工单数据-售后服务投诉 Government Complaints Work Order Data - After-ales Service Complaints），聚焦了政府投诉领域出现频率最高的四个数据场景。它们是由政府投诉部门热线的接线生使用语音-文字转换技术后，进行人工分类标注的。我们在原本庞杂的数据中清洗、纠错、筛选和提炼了四个相关主题来表达真实投诉场景下的文本情况。在数据中我们提供了原始投诉文本和包括诉求类型、诉求分类、客体分类在内的三个标签。数据标签的样本空间可以在一个单独的文件中找到。数据来自于 2021 年 1 月到 2022 年 9 月政府部门接收到的真实文本。

5.2 实验细节

5.2.1 对比实验

在对比试验中，KGCN4Comp 使用了 GCD 数据集中按照投诉场景分层随机抽取的 3450 条（占 GCD 数据集的 80%）已标注的投诉工单数据和 3770 条未标注的工单数据，与异构知识数据融合拼接和知识扩增后，构建了一个节点数量超过 23 万个的知识图谱，用以描述整个投诉处理场景。

在对照组，我们使用了不同编码形式与分类方法的组合，对 KGCN4Comp 在投诉文本分类上的性能进行了对比实验。对照组算法均使用上述混合数据集进行半监督学习，以期在同样的数据训练集情况下进行比较。实验是在 GCD 数据集的不同场景数据上进行的。实验结果可以表明在强基线方法的对比下，我们的方法是否具有更好的分类性能。最终结果如表 3 所示，我们使用 F-measure 参数评估分类器的分类性能，表中数据表明知识图谱在投诉短文本分类任务的知识表达和短文本的理解上有一定的优越性。

表 3 KGCN4Comp 算法与算法组合的 F-measure 与 Accuracy 指标比较

Model	GCD-FSC		GCD-QC		GCD-CC		GCD-ASC	
	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy
Tf-idf+CNN	0.491	0.5171	0.539	0.5634	0.423	0.4376	0.469	0.4992
Tf-idf+R-CNN	0.51	0.5147	0.541	0.5567	0.505	0.5034	0.527	0.546
Tf-idf+BiLSTM	0.52	0.5286	0.57	0.5756	0.472	0.4816	0.501	0.5162
One-hot Encoding+CNN	0.682	0.6879	0.693	0.6966	0.627	0.6271	0.666	0.665
One-hot Encoding+R-CNN	0.687	0.6937	0.667	0.6173	0.637	0.6195	0.622	0.6231
One-hot Encoding+BiLSTM	0.697	0.7042	0.609	0.6103	0.725	0.7338	0.666	0.6571
Bert+CNN	0.84	0.8426	0.834	0.8347	0.826	0.827	0.802	0.8052
Bert+R-CNN	0.847	0.8471	0.839	0.8412	0.827	0.8261	0.816	0.8135
Bert+BiLSTM	0.868	0.8628	0.848	0.8505	0.851	0.8517	0.803	0.8092
KGCN4Comp	**0.893**	0.**893**	**0.852**	0.**8526**	**0.871**	0.**8692**	**0.817**	0.**8173**

5.2.2 消融实验

图神经网络结构如图 3 所示，我们设计了如下的消融实验来判断其每个部分的有效性，并且在适当的调参下获得了比基线更好的算法性能。该网络的详细参数如下：

表 4 KGCN4Comp 各层参数

层名称	参数名称	参数值
自适应特征嵌入层	初始化方法 (Initialization Method)	node2vec
	嵌入特征维度 (Embedded Feature Dimension)	1×512
	激活函数 (Activation Function)	ReLU
	正则化方法 (Regularization Method)	Dropout
	正则化参数 (Regularization Parameter)	0.5
包含多头注意力机制的残差连接层	多头注意力头数 (Number of Attention Heads)	8
	注意力机制隐藏层维度 (Attention Mechanism Hidden Dimension)	64
	残差连接权重 (Residual Connection Weights)	(512, 512)
	初始化方法 (Initialization Method)	零初始化
	激活函数 (Activation Function)	ReLU
	正则化方法 (Regularization Method)	Dropout
	正则化参数 (Regularization Parameter)	0.2
层归一化层	归一化轴 (Normalization Axis)	512
	归一化方法 (Normalization Method)	层归一化
	学习率 (Learning Rate)	0.001
图卷积层	图卷积权重 (Graph Convolution Weights)	维度：(1×512, 1×128)
	激活函数 (Activation Function)	ReLU
	正则化方法 (Regularization Method)	Dropout
	正则化参数 (Regularization Parameter)	0.2
动态池化层	动态池化窗口大小 (Dynamic Pooling Window Size)	7
	动态池化方法 (Dynamic Pooling Method)	最大池化
全连接层	初始化方法 (Initialization Method)	随机初始化
	激活函数 (Activation Function)	ReLU
	正则化方法 (Regularization Method)	Dropout
	正则化参数 (Regularization Parameter)	0.2

在训练中，我们使用 adam 优化器对每次独立实验进行了单独调参，以保证每种算法结构都面临着同样好的训练期望。在 baseline 的选取上我们仅使用了包含图 3 中自适应特征嵌入层和全连接层的简化版本，通过 Adam 优化器对整体模型的参数进行了调优。其他情况是在 baseline 和 ALL 二者的基础上增加或去掉某一层进行的实验，以对比每个层对整个模型的贡献。在以分类准确率为判准的实验中，我们以相同的最好期望分别训练了异构的模型，以期对各个层的贡献得到一个控制变量后的评价。

表 5 显示，相比 baseline，增加任何一层都有 F-measure 指标的提升；而相比 ALL，减少任何一层都有 F-measure 指标的下降。

表 5 KGCN4Comp 的消融实验 F-measure 与 Accuracy 指标比较

Model	GCD-FSC		GCD-QC		GCD-CC		GCD-ASC	
metrics	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy
baseline	0.789	0.7582	0.759	0.7221	0.764	0.7335	0.737	0.7001
ALL	**0.893**	0.**893**	**0.852**	0.**8526**	**0.871**	0.**8692**	**0.817**	0.**8173**
baseline＋MAL	0.798	0.789	0.767	0.7456	0.772	0.7542	0.746	0.7213
baseline＋RCL	0.809	0.8021	0.778	0.7639	0.79	0.7827	0.76	0.7391
baseline＋LNL	0.824	0.8237	0.802	0.7921	0.811	0.8085	0.774	0.7534
ALL-MAL	0.869	0.8673	0.83	0.8268	0.847	0.8436	0.798	0.7947
ALL-RCL	0.875	0.8721	0.836	0.8314	0.856	0.8537	0.803	0.8013

5.2.3 结果和分析

由图 8 可以明显看出，KGCN4Comp 在所有的模型中表现最好。在 GCD 系列数据集的不同场景中，KGCN4Comp 都取得了最高的分类性能。这表明，在这个中文投诉短文本分类任务中，KGCN4Comp 的性能超过了其他所有的模型。

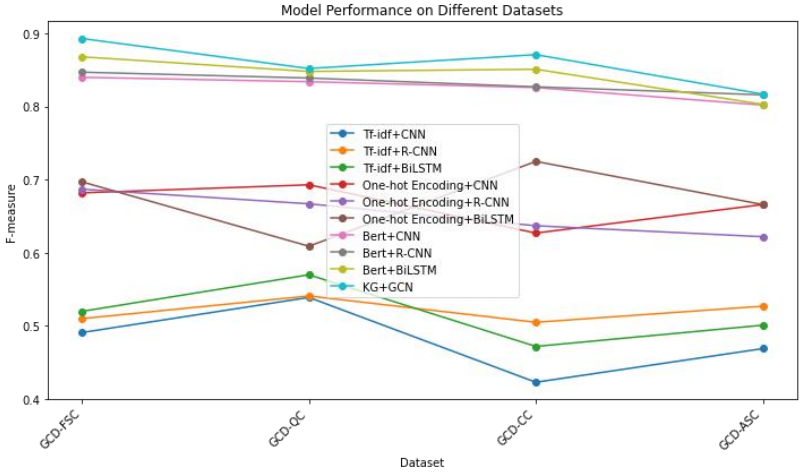

图 8　KGCN4Comp 与其他分类器的 F-measure 性能比较

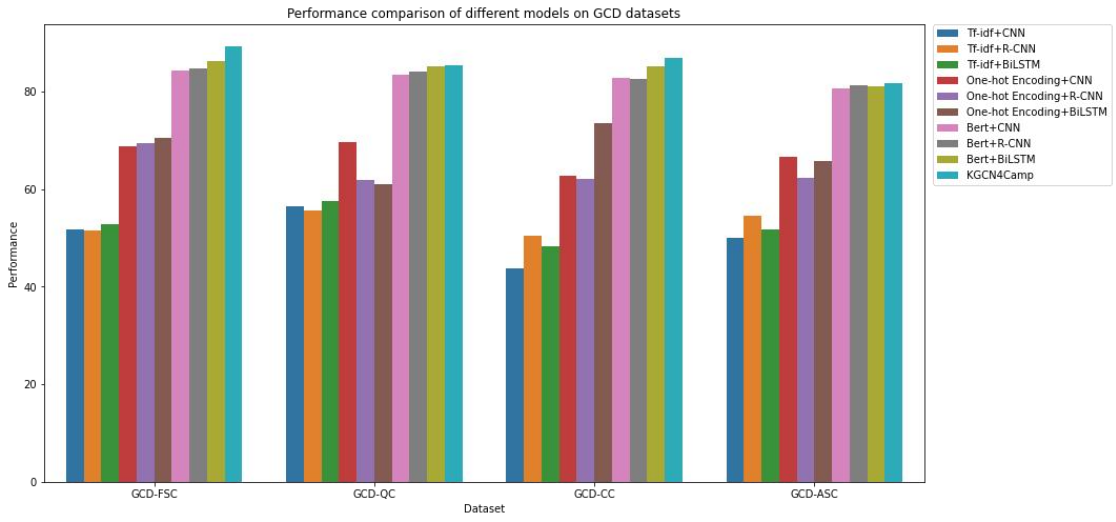

图 9 KGCN4Comp 与其他分类器的准确度比较

最初的 Tf-idf+CNN、Tf-idf+R-CNN 以及 Tf-idf+BiLSTM 模型，其分类准确率相对较低，F-measure 值在 0.5 左右，说明简单的词袋模型和传统的机器学习模型对这个任务的处理能力较弱。随后的 One-hot Encoding+CNN、One-hot Encoding+R-CNN 和 One-hot Encoding+BiLSTM 模型准确率有所提升，但仍然不足以满足需求。这可能是由于这些模型无法充分利用文本中的深层次的语义信息，导致分类能力较差。最后三个基于 Bert 的模型，它们的性能显著提升，说明了深度学习和语义表示能力的优势。特别是 Bert+BiLSTM，在所有的数据集上都取得了非常好的结果。

然而，KGCN4Comp 在所有的数据集上都取得了最好的结果。它在 GCD-FSC，GCD-QC，GCD-CC，GCD-ASC 数据集上的准确率都超过了 85%，F-measure 值在 0.82 以上。这表明，KGCN4Comp 算法能够更好地理解和处理文本数据，特别是对于具有复杂关系和高阶交互的文本数据。同时，图神经网络在处理图结构数据时具有自然的优势，可以有效地抓取和利用图中的结构信息，提升模型的性能。

总的来说，KGCN4Comp 在中文投诉短文本分类任务上的优越性主要体现在以下几个方面：一是能够更好地理解和处理文本数据，特别是对于具有复杂关系和高阶交互的文本数据；二是可以有效地抓取和利用图中的结构信息，提升模型的性能；三是相比其他模型，KGCN4Comp 在所有的数据集上都取得了最好的结果，显示出了更高的准确率和更强的泛化能力。

5.3 问答系统

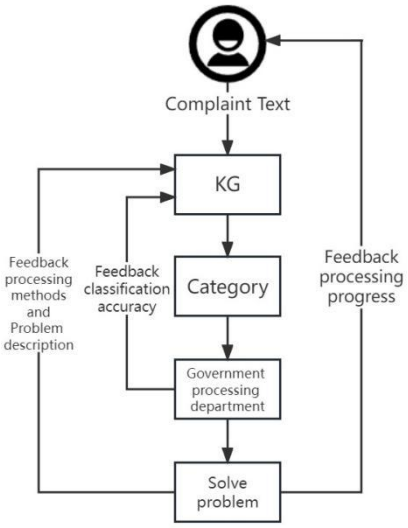

图 10 基于知识图谱的政府投诉平台工作机制

我们利用 GPT3.5 接口实现了一个智能聊天系统，用于处理外来文本投诉。该系统具备自动识别投诉文本的投诉类别、处理部门、处理进度以及预计处理时间等参数的能力，其设计旨在提供一个高效、智能化的解决方案，以满足用户友好性的需求。

在系统的核心部分，我们基于知识图谱技术构建了一个投诉领域的本体模式。该图谱蕴含了丰富的领域知识，包括投诉类别、相关处理部门、处理流程以及预计处理时间等信息。利用这个知识图谱，我们的系统能够智能地将输入的投诉文本与图谱中的实体进行匹配，并确定相应的投诉类别。同时，系统还能够根据投诉类别自动关联适当的处理部门，确保投诉能够被及时转交给相应的责任部门进行处理。我们的系统能够根据投诉情境和用户需求生成相应的回复，提供个性化的支持和解答。图 9 是整个平台的运行逻辑。

六、结论

本研究旨在利用知识图谱和图神经网络的方法，对中文投诉短文本进行多标签分类，以提高分类准确性和效率，构建一个面向政府部门的投诉工单处理机制，提升投诉处理部门对投诉文本的处理能力。此外，在准确率与鲁棒性的原则下，标注和构建了一个应用政府部门官方投诉分类树的基准数据集 GCD。

通过实验验证，我们的方法在中文投诉短文本分类任务上取得了显著的成果。相比传统的文本分类方法，我们的模型在分类准确性上表现出明显的提升，并在收集到的数据集和公开数据集上均得到了较好的效果。

本研究的意义在于探索知识图谱和图神经网络在投诉场景下的中文短文本的多标签分类中的应用潜力，创新性地提出了一个特殊结构的本体模式，增强模型的可解释性。通过利用丰富的知识图谱信息，我们能够更好地捕捉文本之间的语义关系，从而提高分类的准确性和可解释性。

然而，我们的方法仍存在一些局限性。首先，知识图谱的完整性和准确性对分类结果有较大影响，因此根据场景特点进一步改进知识图谱的构建和更新方法是一个重要的研究方向。在实际知识图谱构建中，我们使用了大量来自实际工作中的真实投诉数据，形成了节点规模突破十万级别的图数据参与模型训练。但经由比对我们发现，在政府部门提供的分类方法中，一些分类标签没有被涉及到或只有少量样本分布，长尾效应显著，由此在面对复杂环境的实际场景中，该模型的泛化能力依旧备受挑战。其次，图神经网络的计算复杂度较高，对于大规模数据集的学习和训练可能存在效率问题。

为了进一步提升中文投诉短文本分类的性能，我们建议探索更加有效的知识图谱构建方法，结合其他领域的知识资源进行增强，同时研究如何优化图神经网络的计算效率。在知识图谱中划分基于语义和意群的“联邦”，应用联邦学习的思想综合考虑子图及图边缘节点对整个训练的贡献，并行地计算不同部分也许是一个能够提高模型性能和减少计算时间的思路。

七、引用文献

[1] Luehrmann, L. M. (2003). Facing citizen complaints in China, 1951–1996. Asian Survey, 43(5), 845-866.

[2] Janowski, T. (2015). Digital government evolution: From transformation to contextualization. Government information quarterly, 32(3), 221-236.

[3] Song, G., Ye, Y., Du, X., Huang, X., & Bie, S. (2014). Short text classification: a survey. Journal of multimedia, 9(5).

[4] Yelkenci, B. D., Özdağoğlu, G., & İlter, B. (2023). Online complaint handling: a text analytics-based classification framework. Marketing Intelligence & Planning, (ahead-of-print).

[5] Luo, J., Qiu, Z., Xie, G., Feng, J., Hu, J., & Zhang, X. (2018, October). Research on civic hotline complaint text classification model based on word2vec. In 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC) (pp. 180-1803). IEEE.

[6] Arora, S. (2020). A survey on graph neural networks for knowledge graph completion. arXiv preprint arXiv:2007.12374.

[7] Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017, December). Hdltex: Hierarchical deep learning for text classification. In 2017 16th IEEE international conference on machine learning and applications (ICMLA) (pp. 364-371). IEEE.

[8] Semberecki, P., & Maciejewski, H. (2017, September). Deep learning methods for subject text classification of articles. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 357-360). IEEE.

[9] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1), 273-292.

[10] Zhang, L., Xiao, K., Zhu, H., Liu, C., Yang, J., & Jin, B. (2018, November). Caden: A context-aware deep embedding network for financial opinions mining. In 2018 IEEE International Conference on Data Mining (ICDM) (pp. 757-766). IEEE.

[11] Gao, W., & Huang, H. (2021). A gating context-aware text classification model with BERT and graph convolutional networks. Journal of Intelligent & Fuzzy Systems, 40(3), 4331-4343.

[12] Cekik, R., & Uysal, A. K. (2020). A novel filter feature selection method using rough set for short text data. Expert Systems with Applications, 160, 113691.

[13] Gu, Y., & Shen, J. (2019, November). Short text classification based on keywords extension. In 2019 Chinese Automation Congress (CAC) (pp. 2616-2621). IEEE.

[14] Wang, H., Zhao, M., Xie, X., Li, W., & Guo, M. (2019, May). Knowledge graph convolutional networks for recommender systems. In The world wide web conference (pp. 3307-3313).

[15] Bhowmik, R., & de Melo, G. (2020). Explainable link prediction for emerging entities in knowledge graphs. In The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I 19 (pp. 39-55). Springer International Publishing.

[16] Zhang, J., Chen, B., Zhang, L., Ke, X., & Ding, H. (2021). Neural, symbolic and neural-symbolic reasoning on knowledge graphs. AI Open, 2, 14-35.

[17] Zhang, J., Chang, W. C., Yu, H. F., & Dhillon, I. (2021). Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. Advances in Neural Information Processing Systems, 34, 7267-7280.

[18] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., & Androutsopoulos, I. (2019). Large-scale multi-label text classification on EU legislation. arXiv preprint arXiv:1906.02192.

[19] Jiang, T., Wang, D., Sun, L., Yang, H., Zhao, Z., & Zhuang, F. (2021, May). Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 9, pp. 7987-7994).

[20] Cai, L., Song, Y., Liu, T., & Zhang, K. (2020). A hybrid BERT model that incorporates label semantics via adjustive attention for multi-label text classification. Ieee Access, 8, 152183-152192.

[21] Gong, J., Teng, Z., Teng, Q., Zhang, H., Du, L., Chen, S., ... & Ma, H. (2020). Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification. IEEE Access, 8, 30885-30896.

[22] Ozmen, M., Zhang, H., Wang, P., & Coates, M. (2022, May). Multi-relation message passing for multi-label text classification. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3583-3587). IEEE.

[23] Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. Expert Systems with Applications, 141, 112948.

[24] TIAN Chuang, & ZHAO Yajuan. (2016). A Similarity-based Mapping Model for Patent and Industrial Categories: Taking the International Patent Classification and the National Economic Industry Classification as examples. Library and Information Service, 60(20), 123.

[25] Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(2), 1-49.

[26] Mulang', I. O., Singh, K., Prabhu, C., Nadgeri, A., Hoffart, J., & Lehmann, J. (2020, October). Evaluating the impact of knowledge graph context on entity disambiguation models. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2157-2160).

[27] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24.

[28] Liu, Z., Chen, C., Li, L., Zhou, J., Li, X., Song, L., & Qi, Y. (2019, July). Geniepath: Graph neural networks with adaptive receptive paths. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 4424-4431).

[29] Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

[30] Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.

[31] Jastrzębski, S., Arpit, D., Ballas, N., Verma, V., Che, T., & Bengio, Y. (2017). Residual connections encourage iterative inference. arXiv preprint arXiv:1710.04773.

[32] Xu, J., Sun, X., Zhang, Z., Zhao, G., & Lin, J. (2019). Understanding and improving layer normalization. Advances in Neural Information Processing Systems, 32.

[33] Oishi, Y., & Kaneiwa, K. (2023). Multi-Duplicated Characterization Of Graph Structures Using Information Gain Ratio For Graph Neural Networks. Ieee Access.

[34] Geisberger, R., Sanders, P., & Schultes, D. (2008, January). Better approximation of betweenness centrality. In 2008 Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX) (pp. 90-100). Society for Industrial and Applied Mathematics.

[35] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

[36] Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. Computational Social Networks, 6(1), 1-23.

[37] Jie, H. J., & Wanda, P. (2020). RunPool: A Dynamic Pooling Layer for Convolution Neural Network. Int. J. Comput. Intell. Syst., 13(1), 66-76.