“智慧政务”中的文本挖掘应用

摘要

新一代信息技术的发展使得政府管理更加智慧化，也使得政府需要处理更多文本信息，如公文公报和群众留言等。在这样的情况下，大数据、人工智能技术的发展可以帮助政府人员以新的处理方式应对政务文本数据。基于自然语言处理方法的智慧政务系统以高效简洁的方式提取政务文本中的关键信息。

针对问题一群众留言分类，我们首先对留言主题和留言详情做分词处理，然后引入广义停用词表，并针对问题本身扩展该表，剔除无效文本信息。继而用自然语言处理领域效果较好的 BERT 模型对留言主题和留言详情给予不同的权重，建立一个多分类模型。最后了考虑模型融合的条件及特点，分析了以 BERT 模型为子分类器做模型融合的可行性，创造性的结合了 BERT 模型和 AdaBoost 集成学习方法，并在训练集中得到了准确率与 F1 值的提高。

针对问题二热点问题挖掘，将某一时段内反映特定地点或特定人群问题的留言进行归类。从中可以发现提取出的特定地点或人群要和问题具有一致性，所以我们选择用 BERT-BiLSTM-CRF-NER 模型结合规则统计的方法提取出发生问题频率较高的热点位置或人群 10 个，将其对应的所有留言编号集合记为 A；用规则统计的方法提取出发生频率较高的特定问题 20 个，将其对应的所有留言编号集合记为 B。集合 A 中的每一个子集与集合 B 中的每一个子集取交集，得到新的留言编号的集合。这些新的集合就对应了特定地点或人群的特定问题，并将统计出每一个集合所包含的问题数量作为频数。同时考虑点赞数、反对数和时间跨度作为热度评价指标。采用层次分析法给出每个指标的权数，最终确定热度评价指数，得到排名前五的热点问题。

针对问题三答复质量评价，考虑从相似性、完整性、可解释性和时效性四个角度构建指标体系。由于所有指标无标签，所以借鉴端到端模型的思想，采用多任务学习 XLNet 模型构造三个子任务构造标签。其中前两个子任务的标签采用无监督的方式构造，可解释性标签依赖于人工标注。然后将指标的标签转化为数值型数据。最后通过确定上述指标的权重，运用加法模型得到答复质量评价指数。

关键词：BERT 模型；BERT-BiLSTM-CRF-NER 模型；基于规则的统计；XLNet 模型

Abstract: The development of a new generation of information technology has made government management more intelligent, and it also requires the government to process more text information, such as official bulletins and public messages. Under such circumstances, the development of big data and artificial intelligence technologies can help government personnel respond to government text data in new ways. A smart government affairs system based on natural language processing methods extracts key information in government affairs texts in an efficient and concise manner.

Aiming at problem one, classification of public messages, we first segment the subject and details of the message, then introduce a generalized stop word list, and expand the list for the question itself to eliminate invalid text information. Then, the BERT model with good effect in the field of natural language processing is given different weights to the message subject and message details, and a multi-classification model is established. Finally, considering the conditions and characteristics of model fusion, the feasibility of using BERT model as a sub-classifier for model fusion is analyzed, and the BERT model and AdaBoost integrated learning method are creatively combined, and the accuracy and F1 value are obtained in the training set. improve.

Aiming at problem two, mining hot issues, categorize the messages that reflect the problems of specific locations or specific groups of people within a certain period of time. It can be found that the extracted specific places or people have to be consistent with the problem, so we choose to use the BERT-BiLSTM-CRF-NER model combined with rule statistics to extract 10 hot spots or people with high frequency of problems. The corresponding set of all message numbers is denoted as A; the rule-statistic method is used to extract 20 specific problems with a high frequency of occurrence, and the corresponding set of all message numbers is denoted as B. Each subset in set A takes an intersection with each subset in set B to obtain a new set of message numbers. These new sets correspond to specific problems of specific locations or groups of people, and the number of questions contained in each set is counted as the frequency. At the same time, the number of likes, oppositions and time spans are considered as the

evaluation indicators of popularity. AHP is used to give the weight of each indicator, and finally the heat evaluation index is determined to obtain the top five hot issues.

Aiming at question three-evaluation of the quality of responses, consider constructing an indicator system from four perspectives of similarity, completeness, interpretability and timeliness. Since all indicators have no labels, we borrow the idea of an end-to-end model and use the multitask learning XLNet model to construct three subtasks to construct labels. The labels of the first two subtasks are constructed in an unsupervised manner, and interpretable labels rely on manual labeling. Then convert the label of the indicator into numeric data. Finally, by determining the weights of the above indicators, an additive model is used to obtain the response quality evaluation index.

Keywords: BERT model; BERT-BiLSTM-CRF-NER model; Rule-statistic Model; XLNet model

目录

1. 绪论 5
    1.1 研究背景及意义 5
    1.2 研究内容 5
    1.3 符号说明 6
2. 基于 BERT 的群众留言分类 6
    2.1 针对性数据预处理 7
    2.2 Bert 模型理论 9
        2.2.1 输入输出 10
        2.2.2 Bert 模型结构 10
    2.3 Bert 模型搭建 11
        2.3.1 网络结构 12
        2.3.2 数据准备 13
        2.3.3 BERT 模型训练 13
        2.3.4 模型融合 16
3. 基于热点问题挖掘 17
    3.1 数据预处理 18
    3.2 基于 Bert-BiLSTM-CRF-NER 命名实体识别 19
        3.2.1 Bert-BiLSTM-CRF-NER 模型 19
        3.2.2 模型实体提取 22
    3.3 基于规则的命名实体识别 23
    3.4 基于规则+NER 模型的命名实体识别 24
    3.5 构建热度评价指标 25
    3.6 群众留言热度指数计算 27
4. 答复意见评价 29
    4.1 文本相似性评价 29
    4.2 文本完整性评价 31
    4.3 指标标签构建 31
    4.4 答复质量评价模型 32
5. 参考文献 33

1. 绪论

1.1 研究背景及意义

随着新一代信息技术的发展，政府管理和“互联网+”产生了奇妙的化学反应。政府在信息化的基础上拓宽了管理的渠道和范围，与互联网平台相结合形成了智慧政府，从而可以凭借信息化的手段处理政务。经过这两年的电子政务的基础建设后，政府部门之间、政府与人民、政府与企业之间都生成了大量政务信息。然而，传统的人工处理方式在应对庞大的政务信息时往往显得低效缓慢，难以及时解决政务和相关问题，因此需要大数据相关方法处理管理电子政务信息。基于大数据的智慧政务就是将政府涉及的教育、财政、建设等海量政务信息在一定的模型与工具的处理下，通过形象直观的方式呈现给政府人员，并帮助形成决策的新型政府服务方式。

智慧政务有多种形式的数据，包括人口统计、财务分析、公文公报、群众留言等，其中的数字数据相对容易处理，而结构化的数据，如文本数据难以用传统方法处理。在这样的情况下，大数据、人工智能技术的发展给与海量的政务文本数据以新的处理方式。在自然语言处理技术的基础上，智慧政务系统帮助政府工作人员以高效简洁的方式管理政务信息，挖掘其中潜在的民意民情和企业动向，协助政府做出更有效的决策，实现智能型政府的转变。

1.2 研究内容

本文主要分为三个部分阐述政府文本挖掘的模型方法与处理结果。第一部分关于群众留言分类的问题，我们主要采用 BERT 模型作为分类器。在对文本进行去停用词、去重和不平衡数据处理之后，采用自己架构的 BERT 模型训练分类器。并使用多轮 AdaBoost 方法得到多个 BERT 模型，将这些模型相互融合，得到最后的总模型。该模型在训练集上的准确率和 F1 值得到了明显的提升。

第二部分关于热点挖掘的问题，BERT-BiLSTM-CRF-NER 模型结合规则统计的方法在命名实体识别上有较好的效果。我们发现特定地点或人群与热点问题需要具有一致性，所以我们在用命名实体识别方法在提取出地点和人物的编号后，与发生频率高的特定问题的编号取交集，得到了对应了特定地点或人群的特定问题的集合，并将每个集合的频数作为热度评价指标。然后从频数、点赞数、反对数和时间跨度四个方面建立指标体系，采用 AHP 方法获取权重，得到热点评价指数。根据热点评价指数，排序后可以找出排名前五的热点问题及其地点人物。

第三部分关于答复质量评价的问题，我们考虑从相似性、完整性、可解释性和时效性四个角度构建指标。通过留言详情和答复意见的编辑距离、最长公共子序列、DSSM-LSTM 神经网络三个指标衡量文本相似度；通过是否包含官方文件或网址引用和有效文本长度衡量文本完整度；将群众留言时间与答复时间的间隔

作为时效性指标。对于相似性、完整性、可解释性指标，我们采用 XLNet 模型多任务构造标签，并将这些指标的标签转化为数值型数据。最后根据确定好的指标计算得到答复质量评价。

1.3 符号说明

符号	符号解释
L	BERT 模型中 Transformer 的层数
H	BERT 模型输出的位数
A	Multi-Attention 的数量
α_t	基分类器权重
ε_t	BERT 分类模型分类错误率
C_{t-1}	LSTM 模型 t-1 时刻记忆状态
f_t	LSTM 模型 t-1 时刻记忆状态衰减系数
C_t	LSTM 模型 t 时刻记忆状态
i_t	LSTM 模型 t 时刻记忆状态的衰减系数
h_t	LSTM 模型 t 时刻输出
o_t	LSTM 模型 t 时刻模型输出衰减系数
σ	sigmoid 激活函数
W	线性变化
tanh	LSTM 模型输出的激活函数
$\lambda_{\max}$	判断矩阵最大特征根
CI	一致性指标
CR	一致性比例
RI	平均随机一致性指标

2. 基于 BERT 的群众留言分类

本题群众留言分类问题共分为预处理、模型建立、模型融合和分类处理四个阶段。首先将附件 3 数据中的留言主题与留言详解相结合作为输入数据，对输入文本做停用词处理。采用 EDA 方法处理数据类别之间的不平衡问题，同时删除重复性样本。接下来，我们通过设置针对性的网络结构和转化输入数据来建立适用于本题分类任务的 BERT 模型，在训练模型过程中调整参数以获得更好的分类结果。训练好分类模型之后运用 AdaBoost 算法对模型做 7 轮融合提升分类正确率。最终分类模型在训练集评估指标 F1 为 0.965，群众留言分类流程图如 2-1 所示。

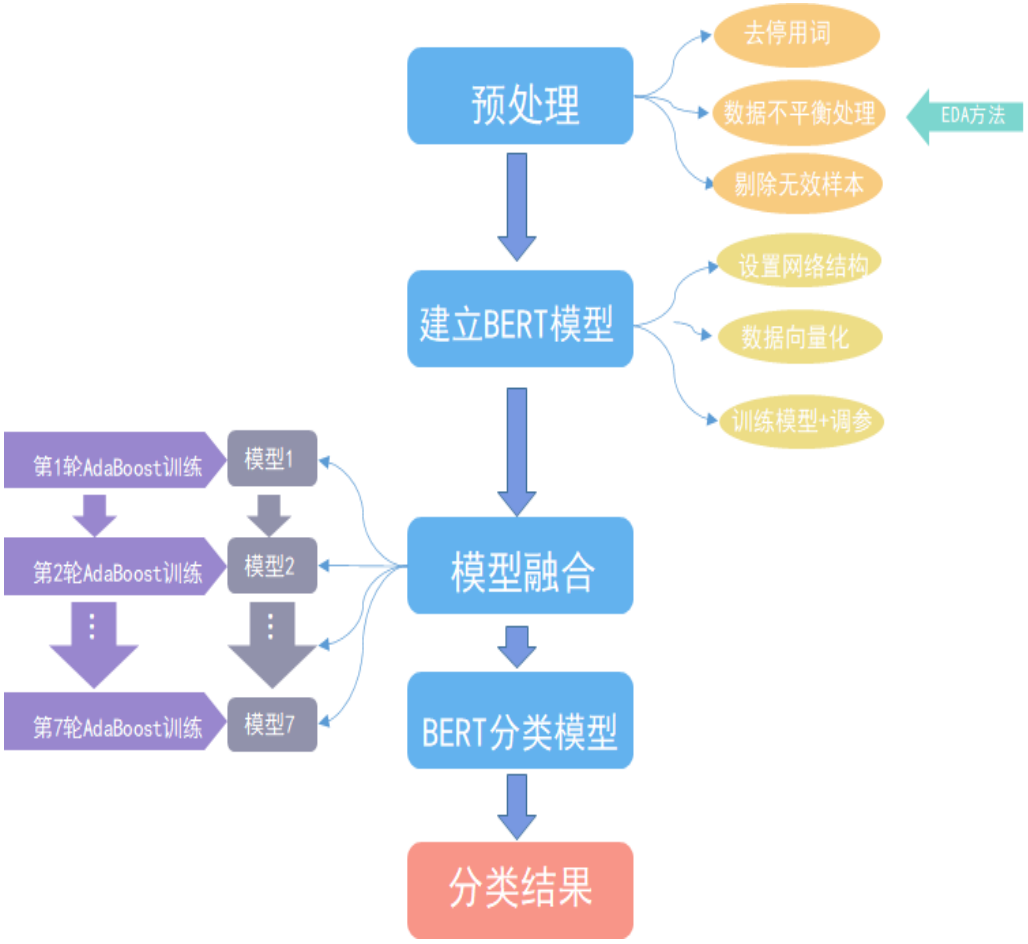

图 2-1　群众留言分类流程图

2.1 针对性数据预处理

1. 去除停用词

停用词是指主动输入的或非自动化生成的，在文本清洗前后需要过滤掉的某些字词。这些词汇高频出现但意义不大，例如：['啊', '哎呀', '哎哟', '唉', '俺', '俺们', '按', '按照']等等，选择将其删去。

在预处理前期，我们利用分词工具将句子切分为多个单词，引入哈工大停用词表，去除停用词，使得输入数据更具有差异性。在训练过程中，根据阶段性结果加入针对该文本任务的停用词。如：在群众留言问题中，“建议”、“问题”、“影响”等高频词汇对于描述民众反应的问题意义不大；“不”等字词虽然对文本语义有影响但是对于问题分类以及判断问题是否是热点也无意义。所以，以上不影响分类效果的词汇在训练中会被剔除。

2. 数据不平衡处理

通过绘制类别标签的直方图，我们可以发现训练集标签类别相对不平衡（图2-2），城乡建设、劳动和社会保障两个类别标签的数量约为 2000 个，交通运输、卫生计生和环境保护类别标签数量均不足 1000 个。

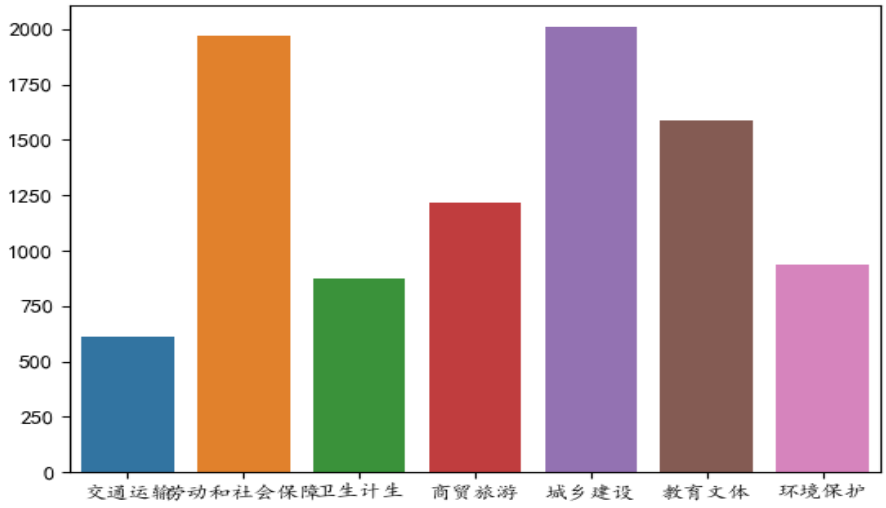

图 2-2 训练集样本标签分布

EDA（Easy Data Augmentation）是一种数据不平衡处理方式，对样本数较少的类别通过同义词替换、随机插入、随机交换、随机删除四种方法生成新的样本。考虑同义词替换法的准确性相对其他三种方法高，所以使同义词替换法在数据增强中占有更高的比例，将同义词替换占有比例设置为 0.2，而其他三种方法的比例设置为 0.1。下表为其中一个处理示例：

表 2-1 EDA 处理示例

原文	处理后
是因为杨辉雪(油籽厂老板)后台大？环保部门受贿真让人弄明白，职权部门无视反贪，保护环境严令请网友转发，请领导严查	杨辉雪公司老板后台环保部门受贿真让人弄明白职权部门无视反贪上诉严令友油籽厂镇严查

由于其他类别生成的新样本与其原有样本的总数不能超过城乡建设类别的数量，所以采用 EDA 的方式将数量较少的三类（交通运输、卫生计生和环境保护）中的留言详情数量分别增加一倍。

我们在训练集（附件 2）中发现，如果用训练出来的高准确率模型去对附件三的无标签数据做一个分类，那么城乡建设类别下的样本数远高于其他类别下的样本数，所以可以认为城乡建设的先验概率在生活中是最高的，应该保持城乡建设类别的高比例样本。测试集（附件 3）样本分布如下图所示：

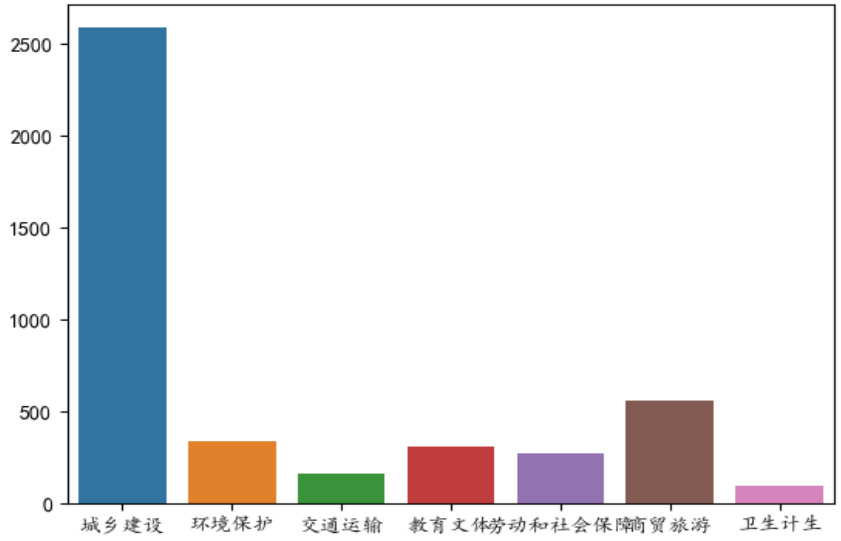

图 2-3 附件 3 分类结果

3. 剔除无效样本

我们发现用户数和留言数不同，所以一个用户可能会有多次留言。在这些留言中，部分内容具有较高的相似性，所以认为同一用户对同一问题的重复留言为无效样本。为了保证样本的真实分布不受无效样本的影响，我们选择剔除这些相似性高的文本。当同一个用户编号下的留言主题有超过 6 个字连续性的相同，那么我们选择剔除这个样本。

2.2 Bert 模型理论

BERT（Bidirectional Encoder Representations from Transformer）是基于 Transformer 的双向编码表示模型，它是 Google 于 2018 年末发布的一种新型语言模型。BERT 模型及其衍生的语言模型在问答、命名实体识别、文本分类等多项自然语言处理任务中发挥着重要作用[1-4]。

对于下游任务存在两种预训练语言表示模型：基于特征和微调的方法。对于基于特征的预训练语言模型，比如 ELMO 模型[1]结合前后词的信息并使用特定的任务框架包括作为附加功能的预训练表示来处理下游任务。基于微调的方法如生成式预训练 Transformer 模型(OpenAI GPT)[5]只考虑前面词的信息，引入最小的任务特定参数，并通过简单地微调所有预训练参数来对下游任务进行培训。这两种方法在训练前都有相同的目标功能，即使用单向的语言模型来学习一般的语言表达。但是单向的语言模型限制了预训练表示的能力，且单向的语言模型在预训练中可以使用的架构类型有限。特别是基于微调的方法，该语言模型仅仅考虑了前面词的信息，比如在 OpenAI GPT 中使用从左到右的体系结构，每个词向量只能关注 transformer 的自我注意层中前一个词向量，但是在句子级任务中需要

结合上下文的信息才能准确把握它的意思。于是提出了基于微调的多层双向Transformer 编码器模型——BERT 模型，该模型受完形填空[6]的启发，在模型结构的预训练阶段增加“遮掩语言模型（MLM）”缓解了单向性的局限。同时基于Transformer 模型不具有时序性，在模型输入阶段也存在着一些改进。自此 BERT 模型从两个方面进行介绍：输入输出和模型结构。

2.2.1 输入输出

BERT 模型的输入数据可以是单词序列中的单个句子或句子对，也可以是连续的文本。对于给定序列，其输入表示可以用三部分嵌入求和组成。嵌入的可视化如下图 2-4 所示：

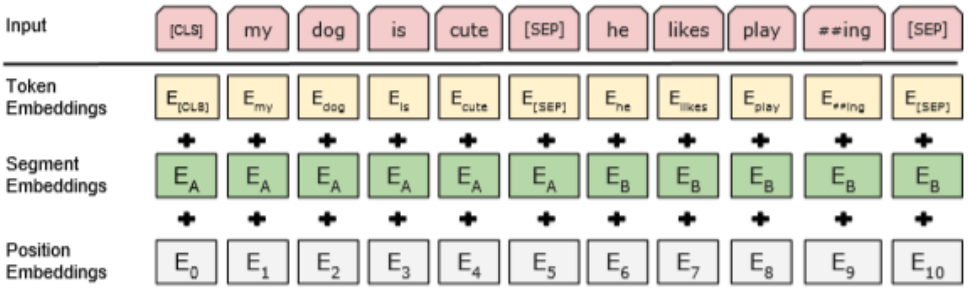

图 2-4　BERT 模型输入

针对中文文本，在词嵌入向量中嵌入 30000 字的字向量表。通过查询向量表将输入序列中的每一个字转换为向量，其中第一个向量是 CLS 标记，可用于后续的分类任务，而对于非分类任务则可以忽略 CLS 标记。Bert 模型能够处理句子对，最后一个字 SEP 作为句子结束的标识。在句子嵌入向量记录输入的文本是输入文本 A 还是文本 B。由于 Transformer 模型不能记住序列的时序，因此 BERT 模型中加入位置嵌入向量用于标记输入序列中每个字出现的顺序。此外，句子的最大长度为 512 位。

针对不同的自然语言处理任务，BERT 模型可以在下游任务的训练中对词向量进行微调，而且模型最后一层的输出也可以有不同的作用。比如，在本题的群众留言分类任务中，BERT 模型在每一个输入文本前插入一个[CLS]符号，并将该符号对应的输出向量作为该用户留言的语义表示进行预测微调。句子嵌入向量中记录句子对中的句子是属于留言主题还是留言详情。

2.2.2 Bert 模型结构

BERT 模型的基础结构如图所示，实际上就是基于 Transformer 结构构造一个多层双向的深度神经网络。Transformer 包括两个独立的机制——读取文本输入的编码器和产生任务预测的解码器。但是 BERT 模型的目标是生成语言模型，因此 BERT 模型又是一个双向的 Transformer 解码机制网络。该模型的创新之处

在于将 Transformer 的双向训练机制应用到语言模型，使得模型可以联合调节所有层的上下文进行预训练。

该模型的结构最初有两个版本，其中 L 表示 Transformer 的层数，H 表示输出的位数，A 表示 Multi-Attention 的数量，同时 Transformer 结构中的前馈神经网络均有 4 层网络结构。BERT 模型结构信息如表 2-2 所示。

表 2-2 Bert 网络结构参数

版本	L	H	A	总参数
BERTBASE	12	768	12	110M
BERTLARGE	24	1024	16	340M

BERT 模型中包含两个阶段：预训练和微调。预训练期间，模型在不同的预训练任务上训练未标记的数据。在微调中使用预训练模型的参数初始化 BERT 模型，然后使用来自下游任务的标记数据对所有参数进行微调，并且每个下游任务都有单的微调模型。BERT 模型的一个显著特点就是在不同任务中具有相同架构，预训练架构与最终下游架构之间的差异很小，且所有参数都经过微调。

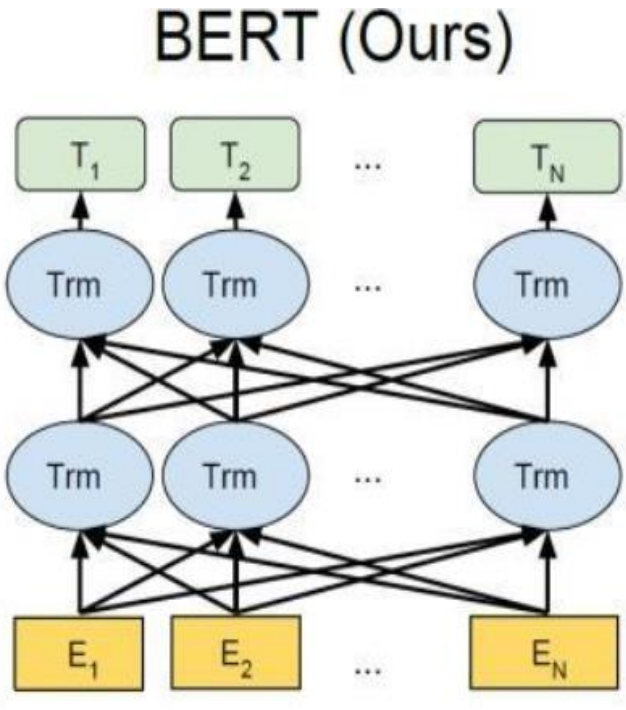

图 2-5 BERT 模型结构

2.3 Bert 模型搭建

由于训练集数据为九千多个，不适合直接训练模型来表示每个字或词的向量，所以本文在做群众留言分类时选用 BERT 模型进行处理。因为稠密向量的分布式训练需要大量的数据，所以首选是使用预训练模型引入先验信息，同时将该问题

作为下游任务。近两年，由 attention 机制衍生出来的 BERT 模型或者 XLNet 模型在各种自然语言处理任务上取得了非常好的效果。虽然两个模型都是双向的，每个词意的向量表示都蕴含上下文的信息，而 Transformer-XL 架构使 XLNet 模型可以学习到更长的句间依赖关系，但是对于该分类任务来说，因为任务足够简单，并不需要学习到长依赖关系，所以选择 BERT 模型作为本题的分类模型。

2.3.1 网络结构

因为数据所包含的信息有限，所以仅将和输出层相连接的全连接层的 dropout 率设为 0.1 防止过拟合。将多头注意力机制数目设置为 12，用以使得句子之间每个字能够相互学习到多样性特征。选择 GELU 函数作为全模型的激活函数，GELU 激活函数的最大特点是将非线性与依赖输入数据分布的随机正则化器相结合在一个激活函数的表达中，GELU 函数和 ELU 以及 RELU 的对比如下图 2-6 所示：

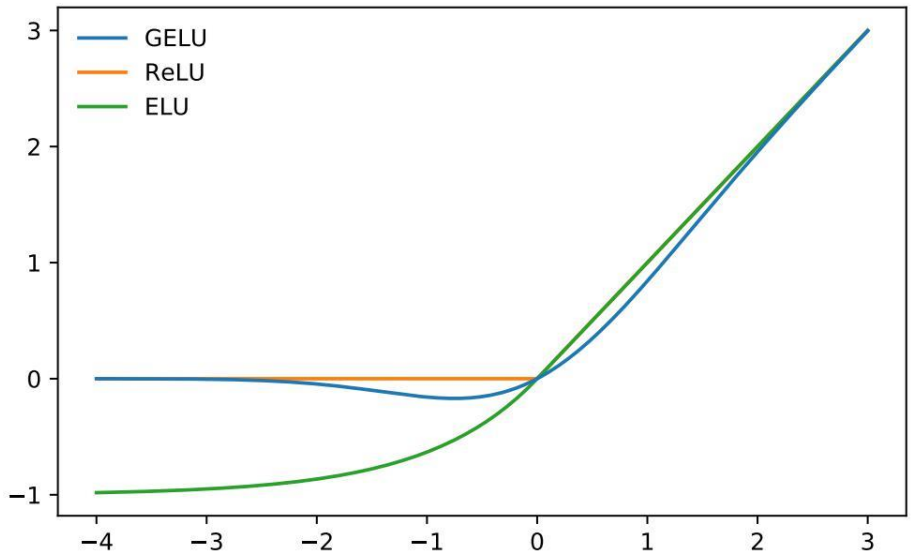

图 2-6 激活函数对比图

选择 GELU 损失函数的原因是是它在 NLP 领域的当前表现最佳，尤其在 Transformer 模型中表现最好。BERT 模型原来的架构已经是 12 层的 encoder 端堆叠，比较复杂，所以只需要加上一个 dropout 层防止过拟合，最后用全连接层即可做分类任务。将模型训练的轮数设置为 7，这么做有以下几个优点：一是减少训练时间并且充分利用数据；二是当模型训练的不够充分看起来是一个弱分类器的时候，正好满足了模型融合的条件；三是随机梯度下降的训练方法加上训练的不充分可以使子模型满足多样性的条件。BERT 微调模型架构图 2-7 所示：

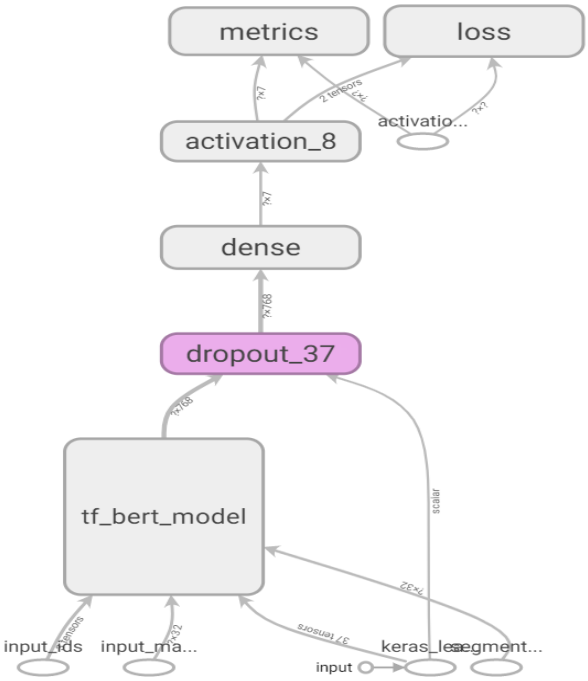

图 2-7 基于群众留言分类的 BERT 网络架构图

2.3.2 数据准备

在输入文本数据时，将留言主题和留言详情作为句子对输入到模型中。为了使得句子长度相同，将句子对总长度设置为 128，长度总和不够 128 的用 0 填补并用 mask 向量标记填补位置，使其不参与训练。为了区分留言主题和留言详情，在 segment 向量中用 0 标记留言主题，用 1 标记留言详情，这样会给予两者不同的权重处理。输入数据示例如下所示：

表 2-3 输入数据示例

原数据	text_a(留言主题)	text_b（留言详情）
	A 市……严重	位于书院路……牵头处理。
Input_ids	101[CLS], 143, 2356,…, 698, 7028, 102[SEP]	855, 754, 741, 7368,…, 1928, 1905, 4415, 511, 102[SEP], …0, 0, 0, 0, 0
Input_mask	1, 1, 1, ……, 1, 1, 1	1, 1, 1, 1, 1,……, 1,1, 1,0, 0, ……, 0, 0
Segments_ids	0, 0, 0, ……0, 0, 0,	1, 1, 1, 1, 1,…… ,1, 1,1, 0, 0, ……0, 0

2.3.3 BERT 模型训练

首先加载在大规模语料库上的预训练模型，使每个字都具有一个基本的向量表达，然后训练时选择 Adam 优化算法和交叉熵损失函数编译模型。由于 TFRecord 只需要一次性加载一个二进制文件的方式即可，所以为了更快的读取

数据在训练前将数据读入 TFRecord 格式。然后使用 tf.data.Dataset 高级 API 将数据送入模型训练，用以计算梯度。由于网络结构较为复杂，很容易产生梯度爆炸的情况，所以这里进行梯度剪切。

1. 训练流程

TFRecord 读入数据之后使用 Iterator 遍历数据，Map 函数将二进制文件解码为模型可以载入的数据。由于一般情况下模型可能会以数据输入的顺序判定类别，所以使用 Shuffle 函数打乱数据集，同时为了避免训练过程中数据缺失情况对数据集做 Repeat 重复。因为模型采用了小批量随机梯度下降的训练方法，所以当某一次迭代剩下的数据量不够一个 batch_size（64）时，对该数据做 drop_remainder 即将数据舍去便于训练方便。由于采用随机化选择数据的方法，被舍弃的数据在后面的训练中有大概率被再次训练，因此并不会影响模型精度。数据流程图如 2-8 和 2-9 所示

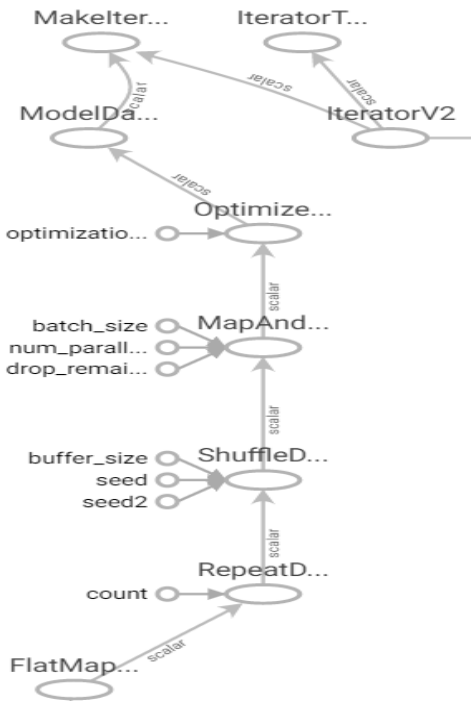

图 2-8 数据流图一

如图 2-9 所示，读入数据让模型作正向传播，图中 loss 节点计算损失函数，反向传播求出每一个可训练参数的梯度。global_norm 可以计算多个张量的全局范数，clip_by_global_norm 函数用作梯度剪切，控制梯度在一个合适的范围防止梯度爆炸的问题出现，继而将变量加以更新来减小交叉熵损失并提高模型分类正确率。

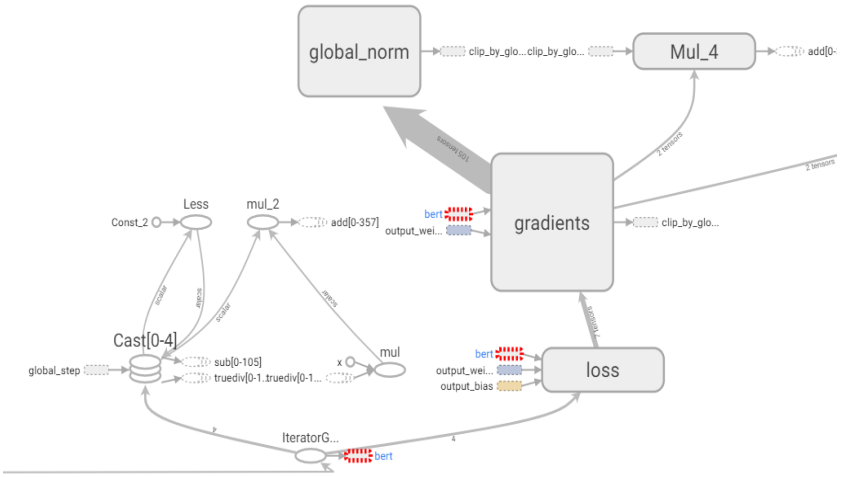

图 2-9 数据流程图二

2.主要参数变化

由于预训练模型本身具有特殊含义，所以在训练中我们对下游任务添加的全连接层的两个变量加以追踪，分别是 kernel 权重变量，和 bias 偏差变量，两个变量随着训练轮数的增加，变量分布变化如图 2-10 所示：

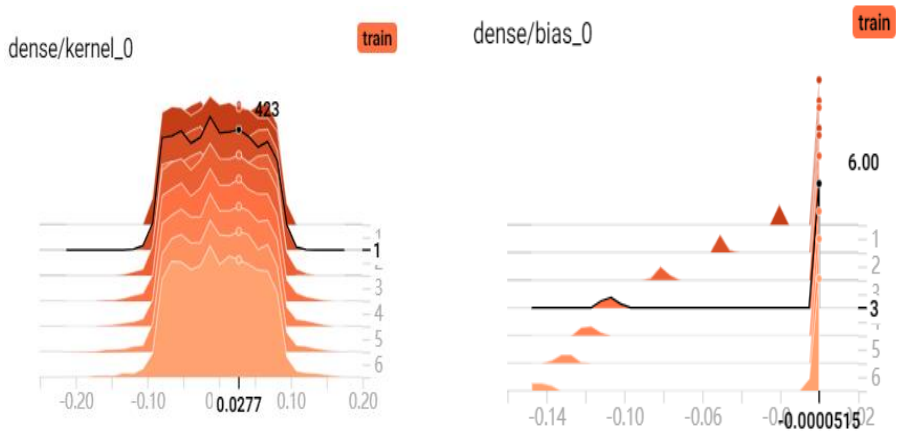

图 2-10 变量分布变化图与权重分布变化图

我们在图中可以看到，全连接层的 kernel 权重变量在-0.1 和 0.1 之间呈均匀分布，0 附近的参数变化不大，这也验证了一点：模型将不太重要的变量学习到了一个较稳定的数值，而变化主要集中于-0.1 和 0.1 附近的变量，逐渐从短尾分布到长尾分布。而 bias 变量则变化较大，说明变动偏差就可以对交叉熵损失产生较大的影响，从这里可以看出偏差依然呈现较大的变化趋势，从这里继续增加训练轮数依然会减小交叉熵损失，但这不一定是有必要的，因为它容易造成过拟合，且和模型融合的特点相违背。所以训练到这里比较合适，接下来还要看模型融合

的效果做进一步判断。

3.BERT 合理性分析

加载预训练模型中最重要的是 embedding 层的参数，对于中文来说，每个字对应一个 768 维的向量，在经过下游任务监督训练的微调之后，对 embedding 层用 t-SNE(t-distributed stochastic neighbor embedding)方法降到 3 维进行可视化分析。我们固定住三维坐标轴对空间中聚集的每一个簇进行标签分析，可以看到，聚集在图中每一个簇的部分分别是第一行（日文，中文生僻字，标点符号），第二行（数字，中文常用字，中文常用字），可以看到 BERT 比较好的用向量来代表一个字符的含义。

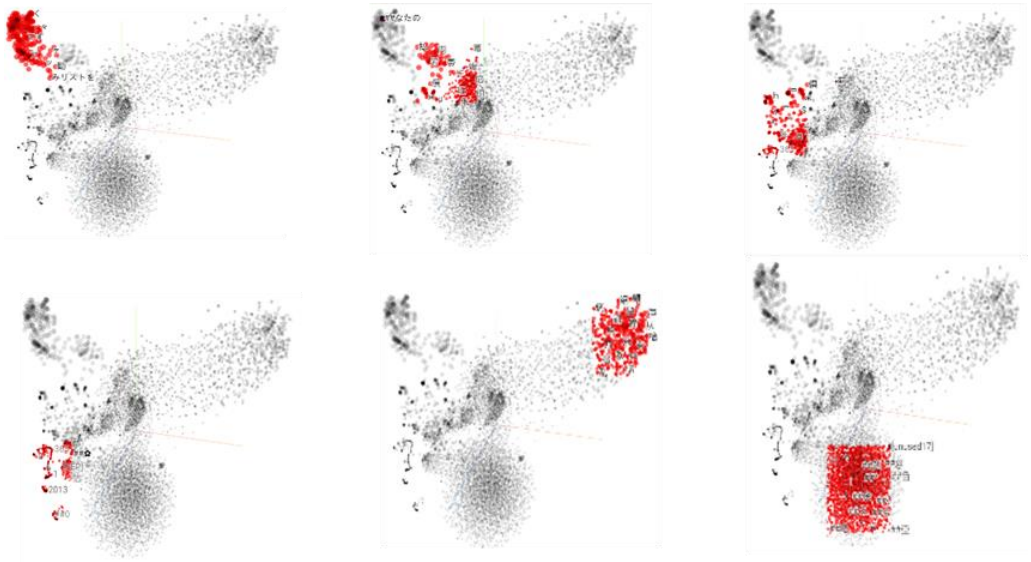

图 2-11 字嵌入向量可视化

2.3.4 模型融合

如果每个子模型可以学习到不同的特征，并且分类正确率都比较高，那么模型融合一定会使模型的效果更好。首先第一次训练 BERT 神经网络，得到模型参数 w 和错误率 ε_t，然后得到权数 α_t 用以调整下一次模型训练中每个样本的分布，使被分错的样本在损失函数中有一个较高的权数，然后循环做此步骤 M 轮。

$$\alpha_t = \frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$$

选择 AdaBoost 串行的模型融合方法，计算已有模型的错误率，算出下一次模型训练中每个正确样本和错误样本的权重进行训练，在第一轮 AdaBoost 训练后，得到模型 1；第二轮训练后，得到模型 2，……，经过 7 轮训练，我们发现其准确率稳定在 97%左右。因此停止继续训练，得到最后的合成模型，其准确率达到了 98%。

表 2-4 模型融合准确率

模型	准确率
模型 1	0.9752416114670431
模型 2	0.9687262460636334
……	……
模型 7	0.9756451346424564
模型融合	0.9833858182213052

3. 基于热点问题挖掘

本题中热点问题是指某一时段内群众集中反映的某一问题。将某一时段内反映特定地点或特定人群问题的留言进行归类后，我们发现提取出的特定地点或人群与热点问题具有一致性，所以选择命名实体识别的方法提取出热门的 10 个含有特点地点或人群的留言编号。我们采用基于规则的方法和 BERT-BiLSTM-CRF 模型作为本题命名实体识别的方法，通过两个方法交叉提取出 20 个特定问题的留言编号。然后将每个地点人群的留言编号与每个特定问题的留言编号取交集，得到 200 个留言编号的集合，其中的每一个集合对应了特定地点或人群的特定问题，最后统计出每一个集合所包含的问题数量作为频数。

经过查阅文献和研究，我们认为频数、点赞数、反对数和时间跨度四个指标可以作为评价热点问题的标准。指标权重由层次分析法(AHP)获得。最后通过加法模型得到热度指数，将热度指数排序后获得排名前五的热点问题。流程图如下所示：

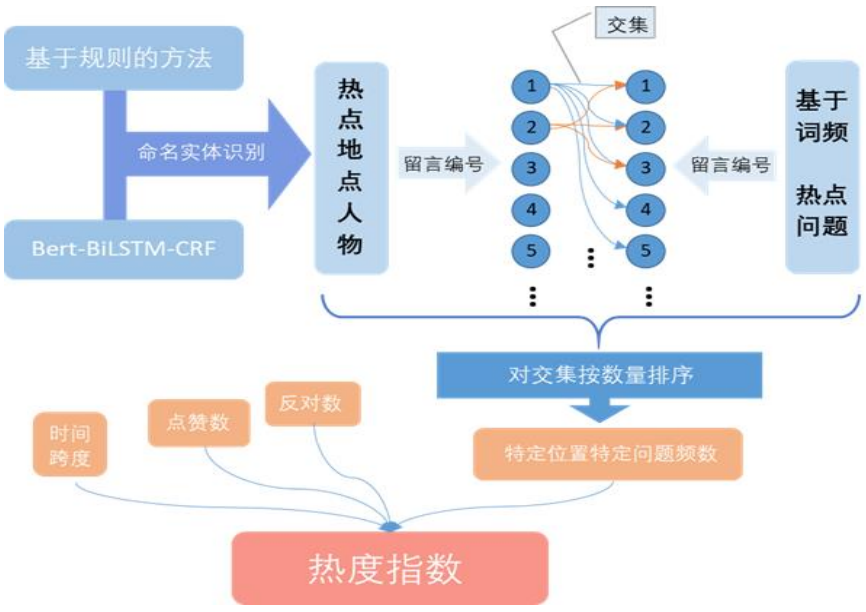

图 3-1 热点问题流程图

3.1 数据预处理

（1）重复性用户发言

网络问政平台的群众留言中部分夹带着私人情绪，会出现同一用户针对同一主题连续发多条相似性留言的情况。这种情况会对热点问题的挖掘带来较大噪声，甚至影响最后热点问题评价的准确性，因此我们需要考虑剔除同一用户的相似性留言。

此外，考虑到留言数量比较多，而普通算法的复杂度是$O(n^2)$，如果用户数量有一万的数量级，那么计算次数则是上亿的。所以这里根据算法本身的特点做了以下优化：对每个用户的留言用递归算法，遍历每一条留言，判断该留言和其余留言是否有相似的留言主题。如果有相似内容，则标记该重复样本，从而不用对该样本做相似性检测。这样保证了对于需要遍历每个用户的留言方法而言，其算法减少了复杂度，即只有$O(n)$。

（2）针对性去停用词

网络问政平台中部分留言的口语化比较严重，存在滥用标点、相同地点叙述不一、错别字等现象。同时这些词汇对热点挖掘没有任何意义，因此在预处理阶段可以将一些无意义的词（比如[“生气”, “!”]和一些不礼貌用语）加入到停用词表，在热点分析时去除相关的词。

（3）点赞数、反对数指标的离散处理

根据图 3-2，点赞数为 0 的样本占据所有样本的 70%，点赞数在 10 以内的样本数占所有样本的 97%。所以极端数据，例如点赞数为 2097 这类只有一个样本的值可能在评价体系中过多的影响最终计算结果。因此我们将它们进行离散化处理，使其在一定区间上分布均匀。

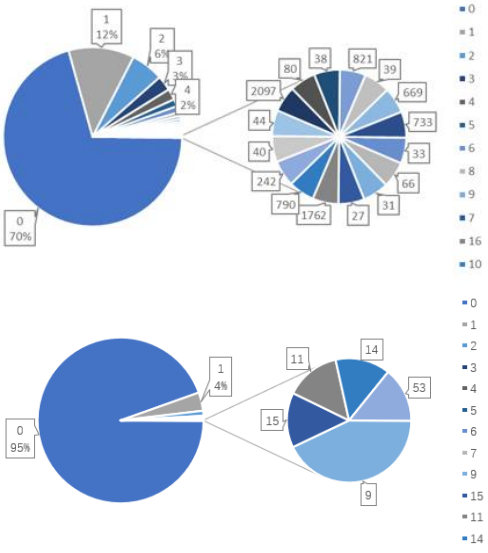

图 3-2　点赞数与反对数频数分布图

与点赞数相似，反对数的为 0 的样本占据总样本的绝大多数，达到了 95%左右，因此也需要做离散化处理，划分一定区间。点赞数和反对数离散后分布如表 3-1 和表 3-2 所示：

表 3-1 点赞数分布表

点赞数	0	1	2	3	4	5	6	7～13	≥13
样本数	3049	514	255	117	86	57	49	105	94
等级	0	1	2	3	4	5	6	7	8

表 3-2 反对数分布表

反对数	0	1	2	3	≥4
样本数	4087	155	42	13	29
等级	0	1	2	3	4

通过将点赞数，反对数两个指标下的值在某个区间或点上均匀分布，我们人为地将其分为上表的区间或值。对于点赞数的后两个区间，我们认为点赞数的数值可以按照等级划分，将点赞数的值处于[7,12]的设置为等级 7，点赞数大于 12 的样本设置为等级 8。对于反对数大于 4 的样本设置为等级 4。

3.2 基于 Bert-BiLSTM-CRF-NER 命名实体识别

命名实体识别（Named Entity Recognition，简称 NER）是信息提取的一个子任务，旨在将文本中的命名实体定位并分类未预先定义的类别，比如人名、位置、组织、时间等等。即识别自然文本中的实体指称的边界和类别。NER 是自然语言处理中一种常见的处理任务，它是一种序列标注问题，因此在实际问题中的数据标注方式也遵照序列标注问题的方式，主要有 BIO 和 BIOES 两种。在 BIOES 标注方式中，B 表示识别某个实体名称的第一个字，I 表示实体中间的字，E 表示该实体最后一个字，S 表示单个字符，O 用于标注与实体无关的字符。

本文分别基于双向 LSTM-CRF 法以及规则法训练 NER 任务抽取留言中的地点或人群，然后根据两种方法的优缺点，采取互补方式将两种方法提取结果结合得到群众留言的实体。

3.2.1 Bert-BiLSTM-CRF-NER 模型

Bert-BiLSTM-CRF-NER 模型中 Bert 主要用于实体识别任务中语言预处理，相当于文本预处理阶段，将文本数据转换为后面命名实体识别模型能够识别的数据类型。由前面 Bert 模型的介绍可知该模型适用性强且充分利用大量无监督文本数据，将语言学知识隐含地引入到特定的任务中，相比于普通的文本处理方法

比如 Word Embedding 它能够有效的反映文本的原始信息。在命名实体识别任务中采用 BiLSTM-CRF-NER 模型，该模型使用双向 LSTM 模型使得在实体识别任务中可以使用过去和未来的输入特征。而使用 CRF 使得该模型可以使用句子级的标签信息，与之前的命名实体识别系统相比，BiLSTM-CRF-NER 模型相对比较稳健，且对词嵌入的依赖比较少，因此对于实体的抽取效果相对较好。Bert-BiLSTM-CRF-NER 模型网络结构如下所示：

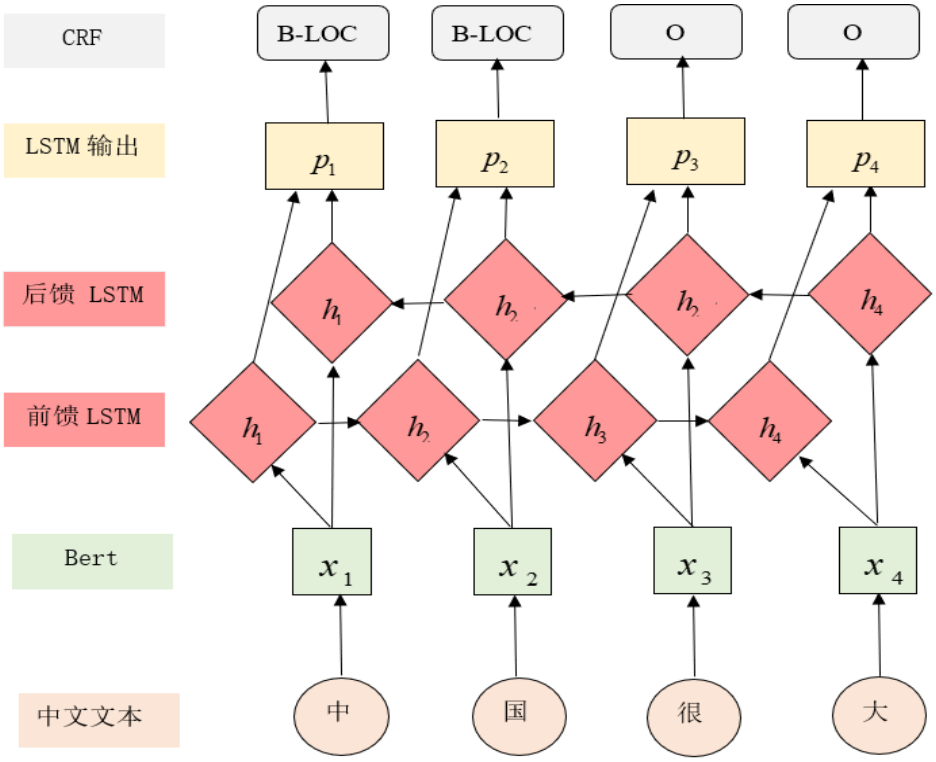

图 3-3 BERT-BiLSTM-CRF-NER 模型网络结构

BiLSTM 网络层是由一个前馈 LSTM 神经网络和一个后馈 LSTM 神经网络组成的双向循环神经网络，LSTM 相比于 RNN 循环神经网络在一定程度上能够解决长时依赖问题，而双向的循环神经网络能够充分的利用文本过去和未来的信息特征，在执行某种任务时准确率更高，稳健性更好。该网络使用时间反向传播（BPTT）来训练双向 LSTM 网络，这样随着时间的推移，双向的向前和向后传递网络与常规网络前向和后向传递的方式相同，不同之处在于我们需要对所有时间步长的隐藏状态设置为 0，还需要在数据的开头和结尾处做特殊处理，该模型可以同时处理多个句子。BiLSTM 的基础架构 LSTM 由三个门来控制，这三个门分别是输入门、遗忘门和输出门。记忆单元是循环神经网络的特殊结构，该结构的存在使得模型具有记忆功能，输出门控制着网络的输出。其中输入门控制网络的输入，遗忘门控制着网络的记忆单元，输出门控制着网络的输出。LSTM 的网络结构由下图所示：

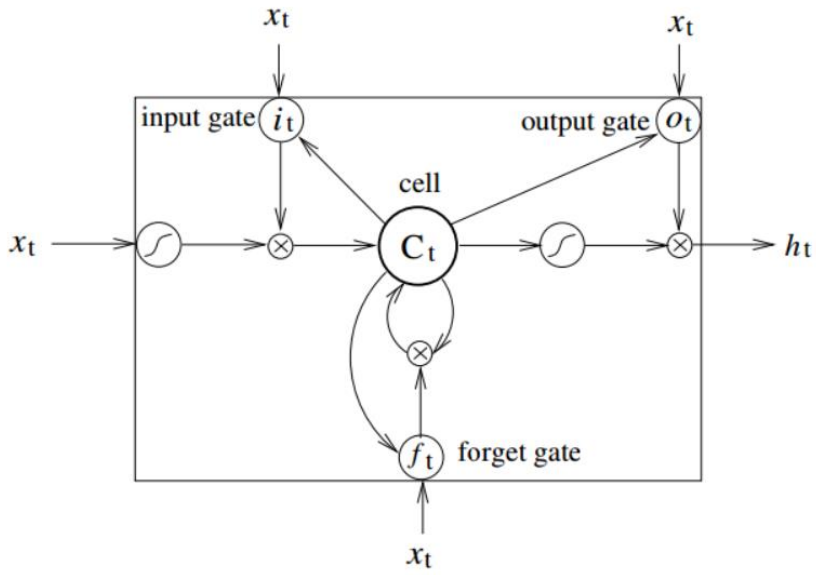

图 3-4　LSTM 的网络结构图

网络结构中 i、f、o 和 c 是和隐藏向量一样大小的输入门、遗忘门、输出门和记忆单元。LSTM 的网络流程是C_{t-1}作为上一步 t-1 时刻网络中的记忆单元，传入 t 时刻的网络之后，LSTM 网络的第一步就是决定上一步记忆单元的遗忘程度，将 t 时刻前面的记忆状态C_{t-1}乘上一个[0,1]的系数进行衰减，其中 t-1 时刻网络记忆的衰减系数f_t是通过 t 时刻网络的输入x_t和 t-1 时刻网络的输出来确定的。接着加上经过学习衰减系数i_t学习到的记忆C_t作为更新之后的记忆传出网络，作为 t+1 时刻网络的记忆单元。其中 t 时刻学习到的记忆根据 t 时刻网络的输入和他 t-1 时刻的网络输出所得到的。

LSTM 网络中的参数具体计算如下所示，其中σ为 sigmoid 激活函数，W 表示线性变化，相当于权重矩阵，W_{hi}是隐藏-输入门矩阵，W_{xo}是输入-输出门矩阵，从记忆单元到输入门的矩阵W_{ci}是对角线矩阵。特别的 t 时刻网络的输出使用 tanh 激活函数。下面是 LSTM 模型的具体公式：

$$\begin{cases} i_t = \sigma\left(W_{x_i} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i\right) \\ f_t = \sigma\left(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f\right) \\ c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc} x_t + W_{hc} h_{t-1} + b_c\right) \\ o_t = \sigma\left(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o\right) \\ h_t = o_t \tanh(c_t) \\ \sigma_z = 1/1 + e^{-z} \end{cases}$$

利用相邻标签信息预测当前标签有两种不同的方法。第一种方法是预测每一步的标记分布，然后使用波束解码来寻找最优的标记序列。最大熵分类器[8]和最大熵马尔可夫模型(MEMMs)[9]的工作属于这一类。第二种方法是关注句子的层

次而不是个体的位置，从而产生输入与输出直接连接的条件随机域(CRF)模型[10]。CRF 是一种无向图概率模型，近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的结果，提取标签的准确率较高。在条件随机场中，给定随机变量 X 条件下，随机变量 Y 的马尔可夫随机场 $P\left(Y\middle|X\right)$，其中 X 为相关的随机变量，例如序列中的具体汉字，Y 为待预测的隐含变量，对应于分词中的 B、M、E、S 等。在线性链中，即 X 和 Y 为一串连续的序列，最大团就是相邻的两项，则条件随机场的计算如下：

$$P\left(Y\middle|X,\lambda\right)=\frac{1}{Z\left(x\right)}\exp\left(\sum_{ji}\lambda_j t_j\left(y_{i-1},y_i,x,i\right)+\sum_{ki}\mu_k s_k\left(y_i,x,i\right)\right)$$

3.2.2 模型实体提取

我们通过 Bert-Base 库打开后台 client 服务端，然后指定模型参数路径和二进制文件路径。输入原文本，得到了每个命名实体的类别边界。利用该模型提取的特定位置或特定人群实体如下表所示：

表 3-3　BERT-BiLSTM-CRF 模型提取结果频数表

地点	频数
伊景园滨河苑	22
丽发新城小区	18
丽发新城	16
星沙	11
A 市伊景园滨河苑	9
……	……
劳动东路魅力之城	6
劳动东路魅力之城小区	4
A 市万科魅力之城商铺	2

模型可以很好的识别出位置信息的边界，经词频统计发现提取的位置信息遗漏并不多，效果很好。但是有如下问题：由于边界不同，模型将“伊景园滨河苑”和“A 市伊景园滨河苑”识别成了两个位置。为此，我们将相似地点进行合并，效果如下表所示：

表 3-4 提取结果合并统计表

地点	频数
A7 县星沙	62
丽发新城	55
伊景园滨河苑	40
A7 县泉塘街道	35
魅力之城	28
A 市地铁	28

3.3 基于规则的命名实体识别

Bert-BiLSTM-CRF-NER 模型模型在命名实体识别上有良好的表现，但是通过结果我们观察到，Bert-BiLSTM-CRF-NER 模型模型对于部分地点，如地铁、公交车等公共场所以及英文数字都难以识别，此外还存在着提取为空值的情况。因此。考虑另一种方法弥补 Bert-BiLSTM-CRF-NER 模型模型的缺陷。在经过研究后，我们发现留言主题具有一定的规律存在，选择基于规则的命名实体识别方法可能有较为良好的效果。基于规则的方法大多通过构造规则模板,选用包括标点符号、关键字、方向词、位置词、中心词等特征，以特定模式和字符串相匹配为主要手段。本文也是利用相应的规则模式提取实体信息。

研究附件三中留言主题可以发现，在大多数留言记录中，其主题主要呈现的是地点/人物+（状语+）动词+（状语+）某事，如某市某小区在夜间……，或者动词+地点/人物+（状语+）动词+（状语+）某事，如请问/反映/咨询某市某地关于/存在/解决……的问题。从上述形式看，可以认为留言主题中第一个名词或英文数字与其后面连续的名词有较强的可能性代表地点/人物，即可以通过一定的语义规则进行命名实体识别。表 3-5 为基于规则的实体提取结果。

表 3-5 基于规则的提取结果

留言主题/划分词性	提取结果
A /市/ A3/区/中海/国际/社区/三期/四期/空地/夜间/施工/噪音/扰民	A 市 A3 区中海国际社区
eng/ n/eng/ n/ ns/ n/ n/ t/ f/ n/ t/ vn/ n/ n	
A2/ 区/ 卢富/ 原著/ 学区/ 划分/不合理	A2 区卢富原著学区
eng/ n/ nr/ n/ n/ v/ n	

根据以上规则，首先需要对短文本进行词性标注，再提取第一个连续性名词组，该连续性名词组就是我们需要识别的实体。我们通过调用 jieba 库对留言主题进行词性标注。由于在文本中存在字母和数字，这些符号同样代表地点或人物，因此除了名词组外，我们将字母和数字纳入提取范围。由表 3-7 结果可知，基于第一个连续性名词组规则的提取结果具有一定的合理性。

表 3-7 提取结果对比表

Bert-BiLSTM-CRF 模型的提取结果	基于规则的提取结果
NaN	**A3 区一米阳光婚纱艺术摄影**
A7 县春华镇金鼎村	A7 县春华镇金鼎村水泥路
黄兴路步行街大古道巷	A2 区黄兴路步行街
中海国际社区	**A 市 A3 区中海国际社区**
保利桐梓坡路麓松路	A 市地铁
NaN	**A 市 6 路公交车**
NaN	**A3 区保利麓谷林语桐梓坡路**
7 县特立路东四	A7 县
……	……

注：加粗的实体相对准确。

通过以上部分提取结果的对比，可以发现基于规则的提取结果对于整体数据有较好的适用性，能够提取相对良好的结果。而 Bert-BiLSTM-CRF-NER 模型模型的结果虽然在部分数据上无法提取任何有效信息，但是比基于规则的方法提取的内容更为准确和详细。因此，两者互为补充形成的结果可以具有比以往更好的结果。

3.4 基于规则+NER 模型的命名实体识别

1.基于规则的方法与 NER 模型的命名实体识别结果的合并

Bert-BiLSTM-CRF-NER 模型和基于规则的命名实体识别方法能够相互补充，因此我们可以将他们的结果相结合，形成新的效果更好的实体识别。

在合并时，由于 Bert-BiLSTM-CRF-NER 模型提取的地点/人物更为精准，所以我们优先采用 Bert-BiLSTM-CRF-NER 模型的结果。对于 Bert-BiLSTM-CRF-NER 模型未提取的结果，直接将规则法的结果作为合并后的结果。对于两个模型都存在结果的样本，我们取两者并集作为最后结果。同时对于规则法中较短的文本提取结果而 Bert-BiLSTM-CRF-NER 模型结果更长、更精确的情形下，我们选择 Bert-BiLSTM-CRF-NER 模型的结果作为提取结果。部分提取结果如表 3-8 所示：

表 3-8 合并的提起结果

Bert-BiLSTM-CRF 模型	基于规则	合并的提取结果
NaN	A3 区一米阳光婚纱艺术摄影	A3 区一米阳光婚纱艺术摄影
A7 县春华镇金鼎村	A7 县春华镇金鼎村水泥路	A7 县春华镇金鼎村
黄兴路步行街大古道巷	A2 区黄兴路步行街	A2 区黄兴路步行街大古道巷
中海国际社区	A 市 A3 区中海国际社区	A 市 A3 区中海国际社区
保利桐梓坡路麓松路	A 市地铁	保利桐梓坡路麓松路
NaN	A 市 6 路公交车	A 市 6 路公交车
NaN	A3 区保利麓谷林语桐梓坡路	A3 区保利麓谷林语桐梓坡路
7 县特立路东四	A7 县	A7 县特立路东四
……	……	……

2. 留言问题归类

通常情况下，一个问题所属类别由动词确定。为此，我们用 jieba 库对留言主题做分词处理，引入动词表，统计留言主题中的动词词频。动词出现频率前二十的词汇如下所示：['扰民','施工','建设','拖欠','改造','销售','经营','生活','办理','诈骗','污染','补贴','管理','增加','培训','开','申请','拆除','反对','调整']。根据这些动词词汇，我们可以将问题归为“扰民问题”，“施工问题”，“建设问题”，“改造问题”，“销售问题”等等，然后抽取出每个问题对应的留言编号。

由于“扰民问题”，“施工问题”等类似问题可能出现在同一留言主题中，所以最后会做去重处理。

3.5 构建热度评价指标

1.确定热度评价指标

在阅读大量文献的基础上，考虑到题目附件所给的数据指标信息和热点问题的相关特点，本文拟从留言内容、受众反应、时间等三个维度选取热度指标。选定内容关键词频数、点赞数、反对数以及同一留言主题时间跨度作为热度评价指标。其中内容关键词频数是上文命名实体识别方法的提取结果与特定热点问题的交集的数量，反映了某一群体或者某一地点在一段时间内某一问题的集中度，是热点问题的主要特征。点赞数与反对数这两个受众反应能够表明看到该条留言的群众对这一事件的态度，若认为确实存在类似情况会点赞，若不同意该事件则会点反对，点赞数与反对数相当于是群众对某一问题的一个侧面表达，同样也能反映问题热点，所以将其纳入评价指标。选取时间跨度这一时间维度指标，主要考

虑到热点问题是与时间密切相关的一个词，它的定义某一时段内群众集中反映的某一问题，所以将该指标加入进去。

2.热度评价指标权重分析

评价指标选取了频数、点赞数、反对数和时间跨度四个指标，为了指标的可解释性和尽可能保留指标的潜在信息，本文不采用缩减或旋转因子的主成分分析法和因子分析法，而采用层次分析法（AHP）实现权重的计算。层次分析法可以通过指标的两两比较获得其在整体数据中的重要程度，是定性与定量方法的结合，能够合理地给出每个决策方案的每个指标的权数[11]。下面是计算权重的具体过程。

（1）构造判断矩阵

判断矩阵是将所有因素两两比较而得到的一种评价矩阵。它通过采用相对尺度来尽可能减少比较不同性质因素的困难。我们发现频数、点赞数、反对数和时间跨度这四个指标自身差异性极大，且个别指标常出现零值（点赞数、反对数），如果仅仅采用按方差贡献值计算的主成分分析法和因子分析法，那么频数的重要性可能低于点赞数、反对数，甚至可能远低于时间跨度。而在现实生活中，热度评价的关键性指标在于相似内容的文本在一定时间内出现的次数，因此可以认为频数明显比其他指标重要，采用判断矩阵衡量指标之间的重要性是合理的。

根据以往的文献研究[12-14]，发现相似文本出现次数的重要性相对较高而点赞数和反对数重要性居次，时间跨度相对而言重要性不高。重要性是两两因素比较得出的介于[1,9]的自然数，因素替换位置则重要性为原值的倒数，如频数相对于点赞数的重要性更大，所以频数：点赞数=5，而点赞数：频数=1/5。因此，构造以下判断矩阵：

表 3-9 判读矩阵

	频数	点赞数	反对数	时间跨度
频数	1	5	7	9
点赞数	1/5	1	3	5
反对数	1/5	1/3	1	3
时间跨度	1/9	1/5	1/3	1

注：单元格内数字表示横向指标相对于纵向指标对整体数据的重要性（重要程度用 1-9 衡量）

（2）一致性检验

指标之间差异性越大越能表现数据的多方面影响。一致性检验就是层次分析法中衡量指标是否相似的检验方法。由于判断矩阵非一致时，有 $\lambda_{max} > n$（λ_{max} 为判断矩阵最大特征根；n 为指标的数量）。所以当 λ_{max} 与 n 的差异越大时，判断矩阵越偏向非一致性。

根据一致性指标 CI 公式：

$$CI=\frac{\lambda_{\max}-n}{n-1}$$

得到 CI 值为 0.058。

用 Satty 模拟 1000 次得到的 RI 表，根据 n=4 得到相应的 RI 值为 0.90。

表 3-10　RI **对应表**

矩阵阶数	3	4	5	6	7	…
R.I.	0.58	0.90	1.12	1.24	1.32	…

由此，通过一致性指标 CI 和平均随机一致性指标 RI 计算一致性比例 CR：结果为 0.064，所以 CR<0.1，认为判断矩阵满足一致性。

（3）计算权重

层次分析法的权重是判断矩阵最大特征根对应的特征向量。通过以下步骤求解：

①按列将矩阵归一化：$b_{ij}=a_{ij}/\sum_{i=1}^{n}a_{ij}$

②按行求和：$\bar{w}=(\bar{w}_1,\bar{w}_2,\bar{w}_3,\bar{w}_4)^T$，$\bar{w}_i=\sum_{j=1}^{n}b_{ij}$

③归一化向量，即权重：$W_i=\bar{w}/\sum_{i=1}^{n}\bar{w}_i$

热度评价指标权重如下所示：

表 3-11 热度评价指标权重

指标	频数	点赞数	反对数	时间跨度
权重	0.6427	0.2083	0.1010	0.0480

3.6 群众留言热度指数计算

1.热度评价指标处理

在热点问题归类后，我们可以发现相隔的天数一半在 300 天到 400 天之间，另一半热点问题的间隔天数在 100 天以内。因此为了直观显示时间间隔的差别，我们将少于 100 天的时间间隔设置为等级 1，而多于 100 天的时间间隔设置为等级 2。

表 3-12 热度指标表

反对数	点赞数	频数	相隔天数	时间间隔
4	42	21	382	2
2	42	20	355	2
1	36	22	327	2
0	9	26	55	1
0	12	24	84	1
17	17	14	51	1
0	0	10	307	2
……	……	……		

2.热度指数计算

在指标及其对应权重确定后，我们可以进行留言热度指数计算。本文选择加法模型，通过计算指标*权重的总和来表示热度的终值。热度指数：

$$head\_index = \sum_{i=1}^{4} w_i x_i$$
$$= freq * w_1 + sup\ port * w_2 + against * w_3 + timeInterval * w_4$$
$$=0.64freq+0.21sup\ port+0.10against+0.05timeInterval$$

其中，freq 表示频数；support 表示点赞数；against 表示反对数；timeInterval 表示时间间隔。

根据热度指数，选择排名在前五的热点问题及其相关内容，如表 3-13 所示：

表 3-13 热度指数排名表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	22.744	2018/11/15 至 2019/12/02	人才购房补贴	人们对于人才购房补贴有大量疑问
2	2	21.899	2019/01/07 至 2019/12/28	西地省大量公司	大量拖欠工资和诈骗问题
3	3	21.833	2019/02/14 至 2020/01/07	A7 县星沙	娱乐场所扰民，旧城改造迟缓
4	4	18.632	2019/07/07 至 2019/08/31	A 市伊景园滨河苑	伊景园滨河苑捆绑销售车位
5	5	17.971	2019/11/02 至 2020/01/25	A 市 A2 区丽发新城	小区附近搅拌站噪音扰民和污染环境

经过排序，可以发现最热点的问题是人才购房补贴问题。其他备受关注的问题还有西地省大量公司的工资拖欠和诈骗问题、A7 县星沙娱乐场所的扰民问题、A 市伊景园滨河苑捆绑销售车位问题以及 A 市 A2 区丽发新城小区噪音问题。

4. 答复意见评价

根据问题描述，我们计划给出评价指标的角度有四个：相关性、完整性、可解释性以及时效性。对于时效性的指标，将群众留言时间与答复时间的间隔作为评价对象。对于相关性、完整性、可解释性的指标，借鉴端到端模型的思想，采用多任务学习构造标签，使用在自然语言处理领域表现优异的 Transformer 架构的有监督模型，子任务设置如下：

无监督子任务一：提取答复中的相似文本（字符级），作为一个分类任务，将其设置为标签“无相似信息”，“有相似信息”，“中立”。其中“中立”表示提取的文本可能是无效文本，无法辨别答复和留言是否相似。

无监督子任务二：提取答复中的长度较短的答复（字符级），作为一个分类任务，将其分类标签“不完整文本”“完整文本”，“中立”。其中“中立”表示一些特定情况，比如一部分答复合理但是不够完整。

有监督子任务三：从解决问题的角度依赖人工将回复质量标签标注为“高质量”，“较高质量”，“一般”，“较敷衍”，“敷衍”。

由于子任务三比较重要，所以对回复质量人工标注是必不可少的环节，这里着重对子任务一和子任务二给出构造方案。

具体构建形式见表 4-1：

表 4-1 答复意见质量评价体系表

	一级指标	二级指标	备注
答复意见质量评价体系	相似性	编辑距离 最长公共子序列 DSSM-LSTM 神经网络	XLNet 模型多任务学习构建对应指标标签
	完整性	是否包含官方文件或引用网址 有效文本长度	
	可解释性	人工标注	
	时效性	群众留言时间与答复时间的间隔	

4.1 文本相似性评价

我们通过三个指标衡量文本相似度：留言详情和答复意见的编辑距离、留言详情和答复意见的最长公共子序列、DSSM-LSTM 神经网络。相比仅用一个指标

测算相似度，三个指标可以给出更好的评价质量。

评价指标 1：留言详情和答复意见的编辑距离。编辑距离(Edit Distance),又称 Levenshtein 距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。通常情况下，编辑距离越小，文本相似度越高，但是这里考虑到答复意见较长的很容易有较大的编辑距离，所以这里要消除答复意见长度对相似性的影响。横坐标表示样本数量，纵坐标表示该样本留言详情和答复意见的编辑距离，如图 4-1（左）所示，纵坐标表示编辑距离与答复意见的长度如图 4-2（右）所示：

图 4-1 编辑距离散点图（左）与加权编辑距离散点图（右）

评价指标 2：留言详情和答复意见的最长公共子序列。一个特定序列的子序列就是将给定序列中零个或多个元素去掉后得到的结果(不改变元素间相对次序)，两个文本的最长公共子序列就是从它们的所有公共子序列中选出长度最长的那一个或几个。通常情况下，最长公共子序列越大，文本相似度越高。横坐标表示样本数量，纵坐标表示该样本留言详情和答复意见的最长公共子序列长度，如图 4-2 所示：

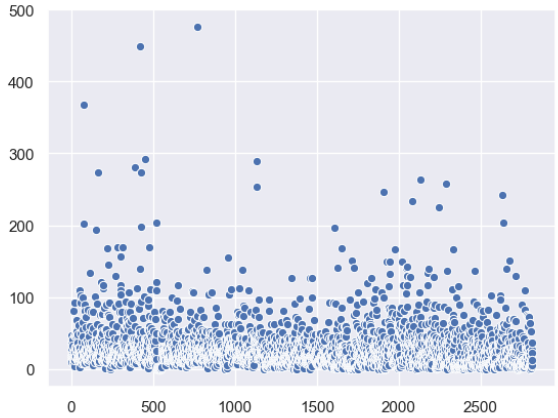

图 4-2 最长公共子序列长度散点图

评价指标 3：用 DSSM-LSTM(Deep Structured Semantic Models-Long Short-Term Memory)计算任意一对短文本的语义相似性。该神经网络能够捕捉上下文信息，适用于留言详情与答复意见这种长度的文本对。虽然它是有监督的学习方法，依赖于人工标注两个文本，但是这种有监督的训练会使模型精度更高。

4.2 文本完整性评价

对于文本完整性评价，我们拟定了两个二级指标：是否包含官方文件或网址引用和有效文本长度。

评价指标 1：是否包含官方文件或网址引用。即如果答复意见中出现了具体文件来对群众提出的问题加以回答，则认为答复意见较为权威，完整，包含设为 1，不包含设为 0。

评价指标 2：有效文本长度。即去除了停用词，并且不考虑“UU0082115 您好获悉”,“感谢您工作关心监督支持”等等文本，横坐标表示样本数量，纵坐标表示答复意见剩余有效的文本长度，如下图所示：

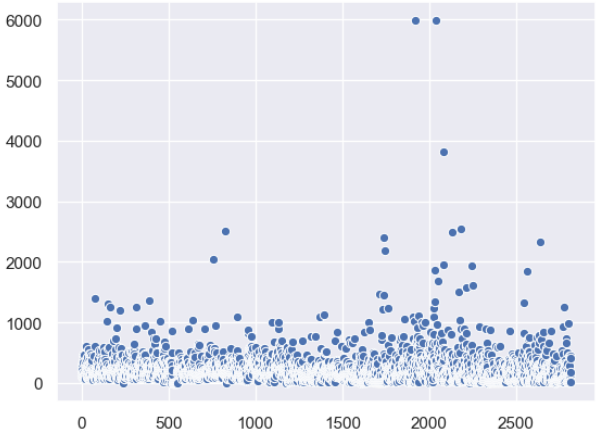

图 4-3 有效文本长度图

可以看到有效文本长度图和编辑距离图有很强的相似性，主要原因是较长的留言回复容易导致更大的编辑距离和有效文本长度。

4.3 指标标签构建

前两个子任务的标签由无监督的方式构造，但是最后一个子任务的标签需要依赖成本高昂的人工标注。

在文本可解释性经过人工标注的情况下，当三个子任务的标签都构造完成之后，选择 Transformer 族模型中的 XLNet 模型，学习答复意见每个句子之间的长依赖关系，按图 4-4 完成模型搭建。模型架构图如下：

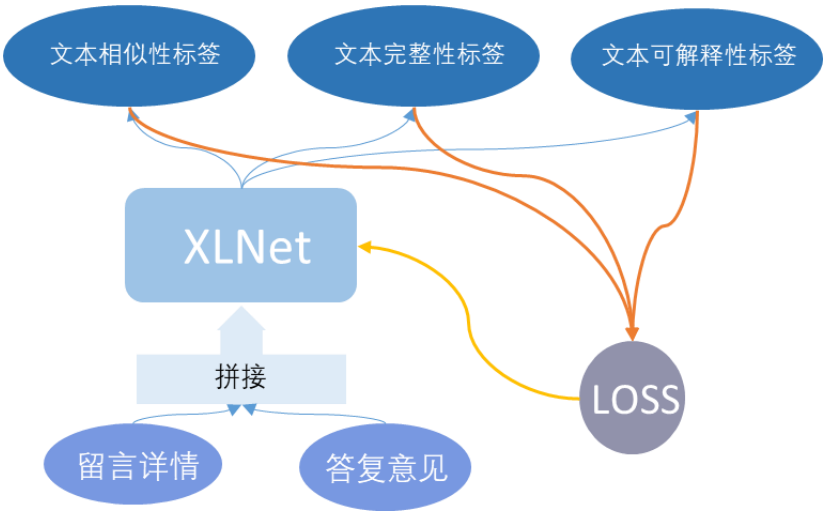

图 4-4 XLNet 模型

通过上图 XLNet 结构，可以得到答复相似性、完整性和可解释性的标签。

4.4 答复质量评价模型

除了题中所给的文本相似性、完整性和可解释性三个角度之外，我们认为答复的时效性也具有较为重要的地位。因此，我们将群众留言时间与答复时间的间隔作为衡量答复是否及时的标准，以便使得答复质量评价更为全面和多样化。

经过以上处理后，我们得到了关于文本相关度、完整度、可解释性以及时效性四个方面的多个指标。我们可以对文本相关度、完整度和可解释性的指标标签数值化处理，从而使得所有指标均变为数值型指标以便于计算。再根据不同的需求，用主成分分析法、因子分析法或是熵权法计算每个指标的权重。最后用乘法模型或加法模型计算评价指数，从而可以衡量每条政府答复的质量。

5. 参考文献

[1] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In NAACL.

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

[3] Andrew M Dai and Quoc V Le.2015. Semi-supervised sequence learning. In Advances in neural information processing systems, pages 3079–3087.

[4] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In ACL. Association for Computational Linguistics.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[6] William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: Better text generation via filling in the . arXiv preprint arXiv:1801.07736.

[7] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

[8] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. Proceedings of EMNLP, 1996.

[9] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. Proceedings of ICML, 2000.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of ICML, 2001.

[11] 张炳江.层次分析法及其应用案例[M].北京:电子工业出版社,2014.

[12]梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.

[13]李小武. 网络舆情热点话题自动发现技术的研究与实现[D].昆明理工大学,2012.

[14]龚海军. 网络热点话题自动发现技术研究[D].华中师范大学,2008.