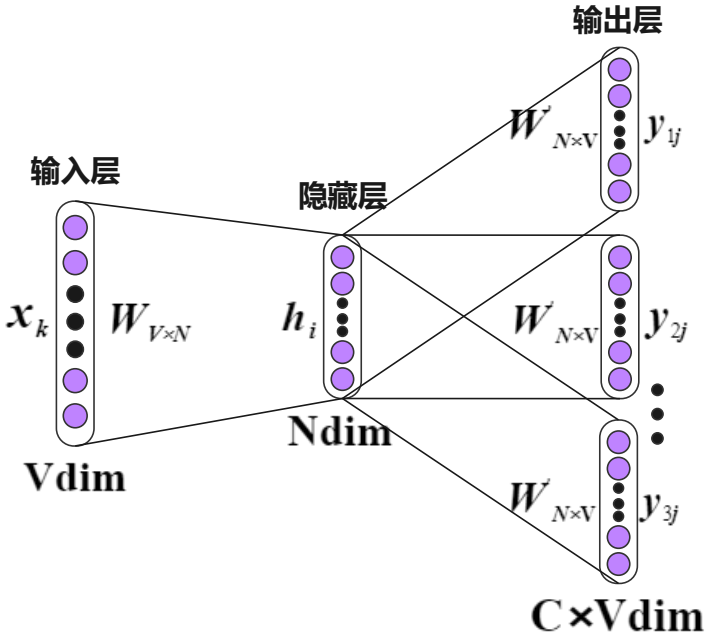
一、实验简介

本实验通过爬虫获取了58634条京东平台手机商品的评价数据，并对其进行了预处理。然后分别构建了基于RNN、LSTM、BERT三种文本情感分类模型，并进行了商品评价的情感分类实战。

二、模型简介

2.1 Word2vec

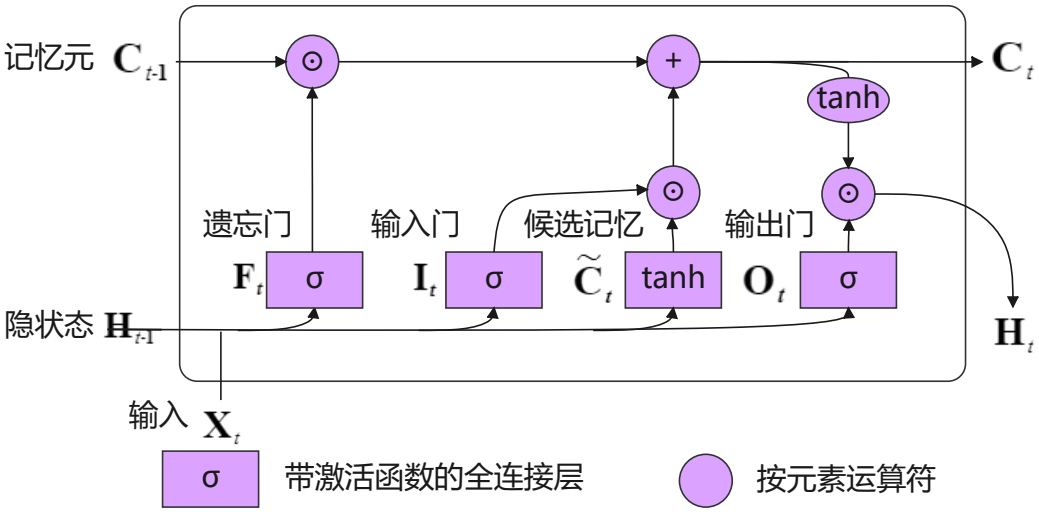
在自然语言处理中，通常将词转化为更具表现力的特征向量。Word2vec通过无监督方式在较短时间内学习到高质量的词向量，相比 One-Hot，其有效避免了维度爆炸且能更好地表达不同词之间的相似性。本实验采用其中的Skip-Gram模型，该模型使用中心词来预估上下文词。Skip-Gram模型如下图所示。



Skip-Gram模型求和梯度的计算成本非常大。负采样是一种类似的训练方法，其通过采样K个不是来自上下文窗口的噪声词，使得每个时间步的梯度计算成本线性依赖于超参数K。

2.2 RNN

循环神经网络常用来处理序列问题。对于时间步t，其通过隐状态保存序列中直到时间步$t-1$的序列信息，其结构展开图如下图所示。



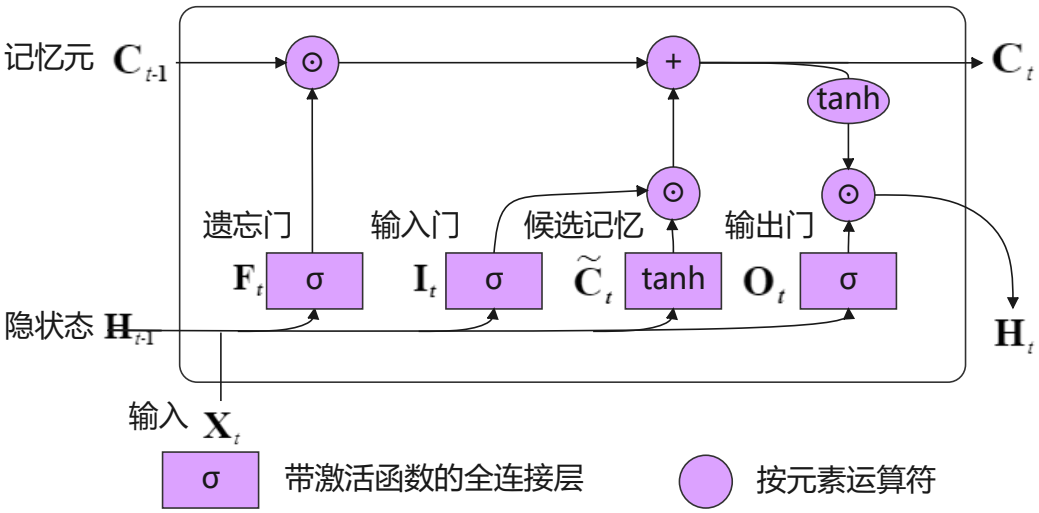
时间步t的隐变量的$\mathbf{H}_t \in \mathbb{R}^{n\times h}$与输出变量$\mathbf{O} \in \mathbb{R}^{n\times q}$的更新函数如下所示：

$$\begin{aligned}\mathbf{H}_t &= \phi\left(\mathbf{X}_t\mathbf{W}_{xh} + \mathbf{H}_{t-1}\mathbf{W}_{hh} + \mathbf{b}_h\right)\\ \mathbf{O}_t &= \mathbf{H}_t\mathbf{W}_{hq} + \mathbf{b}_q\end{aligned} \tag{1}$$

其中，n表示批量大小，d表示输入维度，ϕ为激活函数，$\mathbf{X}_t \in \mathbb{R}^{n\times d}$为输入，$\mathbf{W}_{xh} \in \mathbb{R}^{d\times h}, \mathbf{W}_{hq} \in \mathbb{R}^{h\times q}, \mathbf{W}_{hh} \in \mathbb{R}^{h\times h}$为权重参数，$\mathbf{b}_h \in \mathbb{R}^{h\times h}, \mathbf{b}_q \in \mathbb{R}^{1\times q}$为偏置参数。

2.3 LSTM

随着序列的增加，RNN的反向传播由于过长的矩阵乘积可能会产生梯度消失或梯度爆炸，因此在固定的时间步后需要截断梯度计算，这导致RNN无法有效处理长期依赖。长短期记忆网络增加了记忆元来记录额外信息。模型隐藏层神经元结构如下图所示。



对于时间步t，上图中三个门使用下式计算：

$$\begin{aligned}\mathbf{I}_t &= \sigma\left(\mathbf{X}_t\mathbf{W}_{xi} + \mathbf{H}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_i\right)\\ \mathbf{F}_t &= \sigma\left(\mathbf{X}_t\mathbf{W}_{xf} + \mathbf{H}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_f\right)\\ \mathbf{O}_t &= \sigma\left(\mathbf{X}_t\mathbf{W}_{xo} + \mathbf{H}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_o\right)\end{aligned}$$

其中，$\mathbf{X}_t \in \mathbb{R}^{n\times d}$是输入，$\mathbf{H}_{t-1} \in \mathbb{R}^{n\times h}$为上一时间步的隐状态，$\mathbf{W}_{xi}, \mathbf{W}_{xf}, \mathbf{W}_{xo} \in \mathbb{R}^{d\times h}$和$\mathbf{W}_{hi}, \mathbf{W}_{hf}, \mathbf{W}_{ho} \in \mathbb{R}^{h\times h}$是权重参数，$\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_o \in \mathbb{R}^{1\times h}$是偏置参数。

候选记忆元$\tilde{\mathbf{C}}_t \in \mathbb{R}^{n\times h}$、记忆元$\mathbf{C}_t$与隐状态$\mathbf{H}_t$使用下式计算：

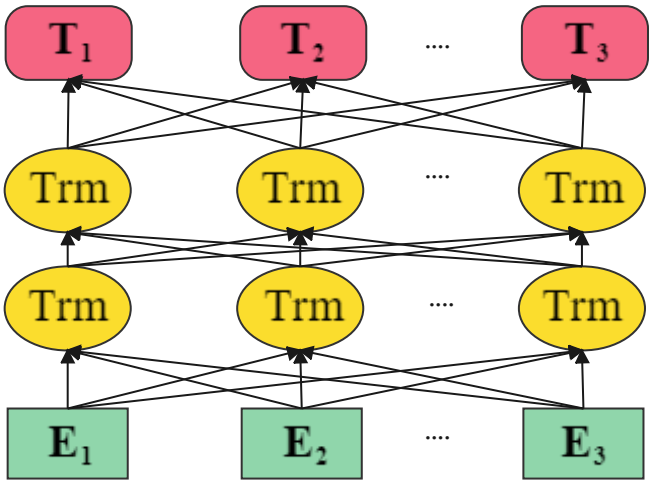
$$
\begin{aligned}
\tilde{\mathbf{C}}_t &= \tanh(\mathbf{X}_t \mathbf{W}_{xc} + \mathbf{H}_{t-1} \mathbf{W}_{hc} + \mathbf{b}_c) \\
\mathbf{C}_t &= \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t \\
\mathbf{H}_t &= \mathbf{O}_t \odot \tanh(\mathbf{C}_t)
\end{aligned} \tag{2}
$$

输入门$\mathbf{I}_t$控制从$\tilde{\mathbf{C}}_t$中获取多少数据，而遗忘门$\mathbf{F}_t$控制从$\mathbf{C}_{t-1}$中获取多少内容。若输出门接近1，就可将所有记忆信息传递给隐状态；若输出门接近0，则不改变隐状态，仅保留记忆元信息。

2.4 BERT

2.4.1 BERT的结构

BERT是一种预训练模型，其采用大量的无标注语料进行无监督预训练，处理下游任务时，仅需附加对应任务的输出层并进行微调即可。这保证了模型能够更加充分的理解语句结构等信息，因此其在下游任务中往往有更好的表现。如下图所示，BERT采用多层Transformer编码器堆叠而成，其中的每个编码器都包含多头自注意力和前馈神经网络。



自注意力在输入数据中增加了位置信息，使模型可以与输入序列中的所有位置进行计算，既可并行计算增加速度也解决了文本的长距离依赖问题。其中的每个注意力头均采用如下的缩放点积注意力。

输入矩阵$X_{embedding}$分别与三个随机初始化的权重矩阵$W_Q, W_K, W_V \in R^{embed*embed}$相乘做线性变换，构成三个与先前维度相同的矩阵$Q, K, V$：

$$
\begin{aligned}
Q &= \text{Linear}(X_{\text{embedding}}) = X_{\text{embedding}} W_Q \\
K &= \text{Linear}(X_{\text{embedding}}) = X_{\text{embedding}} W_K \\
V &= \text{Linear}(X_{\text{embedding}}) = X_{\text{embedding}} W_V
\end{aligned} \tag{3}
$$

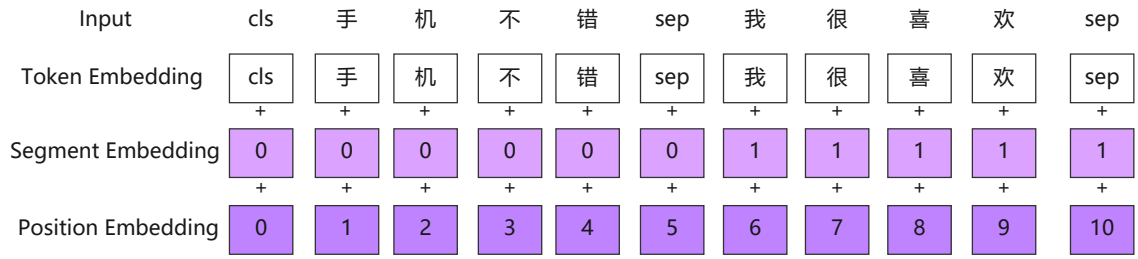
缩放点积注意力计算方法如下所示：

$$
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}
$$

在结果矩阵中, 每一行代表一个字与其他字的相关程度。因此, 由自注意机制所得结果中, 每个字都包含其他所有字的信息。而多头注意力机制就是将 Q, K, V 以 $embedding$ 进 行均分, 使得多个头并行运算。最后将各个头拼接到一起，经过一个全连接层降维输出。

2.4.2 BERT的输入

BERT输入的长度是固定的，通常为512或1024，因此长度不足的位置需要使用""补足。此外，BERT输入都是句子对，""标识经BERT训练后可用于分类任务，""标识用于分开两个输入句子。BERT的输入如下图所示：



- Token Embedding：BERT处理中文任务时以字为单位，无需分词，直接构建词典将输入序列向量化。
- Segment Embedding：区别句子对中的上下两句。通常将上句设为0，下句设为1。
- Position Embedding：记录字在句子中的位置信息。

2.4.3 BERT预训练任务

BERT预训练由两个并行的子任务构成，分别是掩码语言模型（MLM）和下一句预测（NSP）。

为了双向编码上下文以表示每个词元，MLM随机掩蔽词元并使用来自双向上下文的词元以自监督的方式预测掩蔽词元。MLM每次从输入语料中随机选择15%的词元作为预测的掩蔽词元，以“为发烧而生”为例，若被替换的字是“生”，有以下三种情况：

1. 80%的概率被特殊词元更换，即“为发烧而”。
2. 10%的概率更换为随机词元，如“为发烧而存”。
3. 10%的概率无变化，即“为发烧而生”。

NSP以50%的概率将下一句随机替换，50%的概率不发生变化。之后预测两个句子是否真正相邻。

三、模型的构建与实现

3.1 数据采集

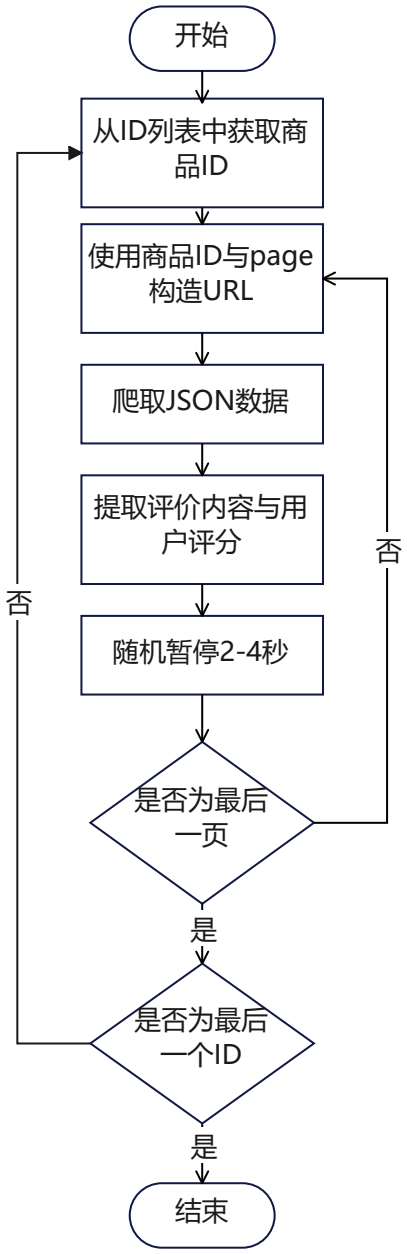
京东商品界面每次刷新评论时，原有链接并不发生变化，其页面会动态加载评论内容。为了获取页面请求的动态内容，打开开发者工具进行数据抓包，并点击评论，可以看到每次刷新评论页面时网页将自动向服务器请求内容。对服务器返回的URL进行分析，可以看到“productId”表示商品编号，“page”表示商品评价的页码等信息。通过更换URL中不同属性，即可实现对JSON数据的获取。某URL返回的JSON数据如下图所示。


```
fetchJSON_comment98({"jwotestProduct":null,"score":0,"comments":[{"id":18537924567,"guid":"9da73b54be2d0f00243a7e10099ea3d3","content":"几天就黑屏了，就变砖了","creationTime":"2022-12-15 10:55:00","isDelete":false,"isTop":false,"userImageUrl":"misc.360buyimg.com/user/myjd-2015/css/i/peisong.jpg","topped":0,"replies":[{"id":1149828302,"commentId":18537924567,"venderId":0,"content":"您好，非常抱歉给您带来不好的购物体验了。使用过程中出现黑屏，可尝试反复多次按动电源键，查看是否可以点亮屏幕；如果锁屏后再次点亮屏幕可以看到锁屏界面，但解锁后依旧黑屏，较大可能为当前正在使用的软件异常导致，可以长按电源键10S强制重启手机。产品使用中有任何疑问您都可以联系客服协助处理，竭诚为您服务。","pin":"小米京东自营旗舰店","userClient":103,"userImage":"misc.360buyimg.com/user/myjd-2015/css/i/peisong.jpg","ip":"120.195.56.79","productId":100017508683,"replyList":[],"nickname":"小米京东自营旗舰店","creationTime":"2022-12-15 11:38:45","parentId":0,"targetId":0,"venderShopInfo":{"id":1000004123,"appName":"//mi.jd.com","title":"小米京东自营旗舰店","venderId":1000004123}}],"replyCount":1,"score":2,"imageStatus":1,"usefulVoteCount":0,"userClient":4,"discussionId":1204506927,"imageCount":0,"anonymousFlag":1,"plusAvailable":203,"mobileVersion":"","videos":[{"id":1992763953,"mainUrl":"https://jvod.300hu.com/img/2022/253192940/1/img5.jpg","videoHeight":960,"videoWidth":720,"videoLength":11,"videoTitle":"","videoUrl":936990466,"videoId":936990466,"stat802d-4278-9a0e-4054b0220143/58f43a0f9731472d902ee6b91b13d1e8.mp4?source=2&h265=h265/113393/b12646e0ae9a4c8ca5f73e60014490c3.mp4"}],"mergeOrderStatus":2,"productColor":"蓝色","productSize":"12GB+256GB","textIntegral":20,"imageIntegral":20,"extMap":{"buyCount":2},"status":1,"referenceId":"100017508683","referenceTime":"2022-12-06 11:18:38","nickname":"****0","replyCount2":1,"userImage":"misc.360buyimg.com/user/myjd-2015/css/i/peisong.jpg","orderId":0,"integral":40,"productSales":"[{\"dim\":3,\"saleName\":\"购买方式\",\"saleValue\":\"标准版\"}]","referenceImage":"jfs/t1/75938/1/17780/65914/62764f02E1e404f6d/500479f21b0bdf31.jpg","referenceName":"小米12 骁龙8 Gen1 黄金手感 6.28英寸视感屏 120Hz高刷 5000万疾速影像 67W快充 12GB+256GB 紫色 5G手机","firstCategory":9987,"secondCategory":653,"thirdCategory":655,"aesPin":null,"days":9,"afterDays":0},{"id":18537809378,"guid":"8a9752a9adf61dfffc363f2aba734bc9","content":"东西已经收到了，物流速度也很快，质量也很好，给好评。","creationTime":"2022-12-15 10:42:49","isDelete":false,"isTop":false,"userImageUrl":"misc.360buyimg.com/user/myjd-2015/css/i/peisong.jpg","topped":0,"replyCount":0,"score":5,"imageStatus":1,"usefulVoteCount":0,"userClient":4,"discussionId":1204498627,"imageCount":1,"anonymousFlag":1,"plusAvailable":201,"mobileVers[{"id":1992750656,"imgUrl":"//img30.360buyimg.com/n0/s128x96_jfs/t1/75336/33/23231/35343/639a89a9E1221938f/dfc39a4f502fbf47.jpg","imgTitle":"","status":0}],"mergeOrderStatus":2,"productColor":"蓝色","productSize":"12GB+256GB","textIntegral":20,"imageIntegral":20,"extMap":{"buyCount":8},"status":1,"referenceId":"100017508683","referenceTime":"2022-12-09 12:49:38","nickname":"韩***诚","replyCount2":0,"userImage":"misc.360buyimg.com/user/myjd-2015/css/i/peisong.jpg","orderId":0,"integral":40,"productSales":"[{\"dim\":3,\"saleName\":\"购买方式\",\"saleValue\":\"标准版\"}]","referenceImage":"jfs/t1/75938/1/17780/65914/62764f02E1e404f6d/500479f21b0bdf31.jpg","referenceName":"小米12 骁龙8 Gen1 黄金手感 6.28英寸视感屏 120Hz高刷 5000万疾速影像 67W快充 12GB+256GB 紫色 5G手机","firstCategory":9987,"secondCategory":653,"thirdCategory":655,"aesPin":null,"days":6,"afterDays":0},
```


可以看到构建获取评价内容爬虫所需的关键属性：

1. maxPage：评论页面的最大数量。
2. content：评论内容。
3. score：用户评分。

在爬取评论内容时，为了防止受到京东服务器的反爬虫限制，在每爬取一页评论后，爬虫将随机停止2-4秒来模拟人为操作。本实验共爬取了58634条数据以用于模型训练的源数据。该爬虫的具体实行步骤如下图所示。



3.2 数据预处理

鉴于在实际购买时，用户通常仅查看商品的好评（好的方面）和差评（差的方面），因此将用户评分为4-5分的评价归为好评，将评分1-3分的评价归为差评。

评价中经常有一些无效信息。本实验对仅包含数字与符号以及长度不足5的评论进行了删除。对于重复评论使用set函数删除。之后，打乱数据顺序来增强随机性，并删除部分数据以保证正向评论和负向评论具有相同数量。最终，将剩余的所有评论以包含文本和评价标签的字典格式分别存放到训练集和测试集对应的文件中，其中训练集占比90%，测试集占比10%。最终使用训练语料条数如下：

1. 正向评论数量:28811（50%）
2. 负向评论数量:28811（50%）
3. 训练集评论数量:51860（90%）
4. 测试集评论数量:5762（10%）

3.3 基于RNN、LSTM的商品评价分析模型

3.3.1 分词与词典构建

词是含有语义信息的基本单位，为了让模型更好的利用语义信息，通常选择将中文分词。本实验采用jieba分词来实现中文分词功能，并通过比对最终选择表现最佳的精确模式。

停用词指自然语言中出现频率非常高但无实际意义的词语，其存在并不影响句子本身的含义，但大量的停用词会影响模型训练以及执行的速度。本实验选择“中文停用词表” 对文本分词后形成的词列表进行匹配过滤。去除停用词的处理效果是显著的，在不去除停用词前，最终所得词典包含10281个词，去除停用词后，所得字典包含10747个词，减少了466个词。

在构建词典时，根据词在语料中出现的频率进行降序排序，并去除出现次数小于5次的词以降低后续模型计算的复杂性，同时添加特殊词元来标记未知词元，特殊词元进行句子填充以统一语句长度。之后对词进行编码，得到其对应的编码映射。

最后。使用构建的词典将所有语料转化为计算机可处理的编码向量。

3.3.2 Word2vec预训练

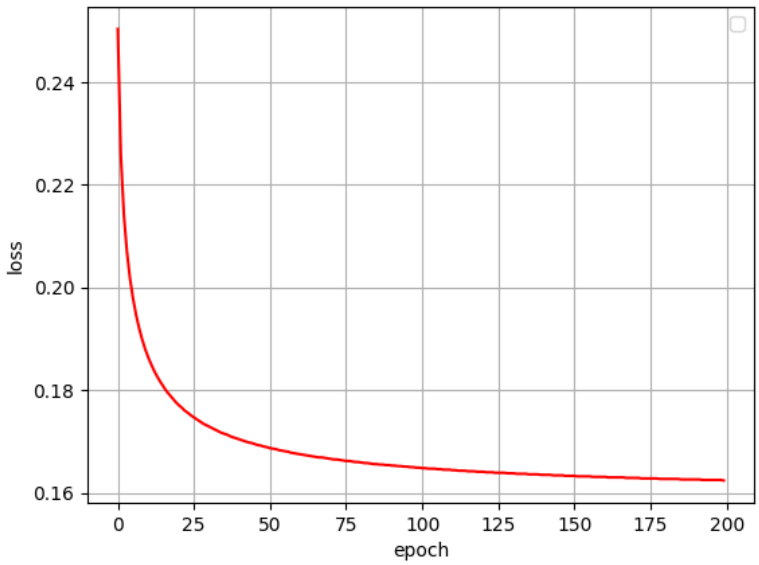
Skip-Gram模型首先从语料中提取所有中心词及其对应上下文词。由于语料分词后词数小于等于5的语句占比达到92.5%，且在设定不同最大窗口参数的对比实验中，参数为2的情况下模型能更好的查找出替换词，因此将最大窗口大小设置为2。对每一个字典化后的向量，对其中的每一个词构建其作为中心词的语料对，其对应上下文词为其左右两侧长度为[1,max_window_size]的向量。

接着将噪声词的采样频率设置为其在字典中的相对频率，其幂为0.75。通过反复实验，最终将超参数K的长度指定为12。对于先前得到的每一个上下文词向量，均以设置的采样频率获取K倍长度的负向量。

然后，以指定的bath_size将语料转换成小批量样本，每个样本包括中心词及其n_i个上下文词和m_i个噪声词，由于每个中心词使用的窗口大小不同，$n_i + m_i$对于不同的i是不同的。因此，在连接上下文词和噪声词时，以$\mathbf{max}(n_i + m_i)$为基准使用进行填充。为了区分正反例，还构建了一个labels向量将上下文词与噪声词分开。

最后，构建具有2个嵌入层的Word2vec模型，使用带掩码的二元交叉熵损失进行模型训练，相应参数如表所示。

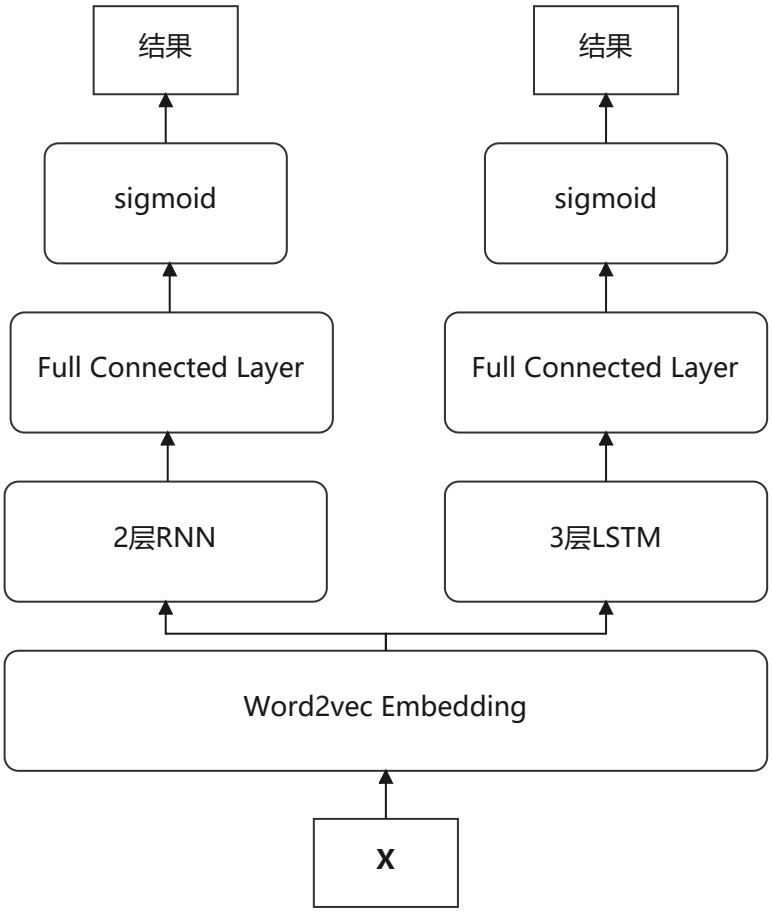
参数名	参数值	参数意义
lr	0.002	学习率
epochs	200	迭代次数
embed_size	100	词向量长度
batch_size	1024	每批运行样本数



模型的嵌入层维度设置为100，这是因为在增加嵌入层维度后模型提升并不明显，因此选择维度更低的模型降低复杂度。

3.3.3 模型构建

针对商品评论是序列文本的特点，本实验构建基于RNN、LSTM的情感分析模型。两个模型的结构如下图所示。

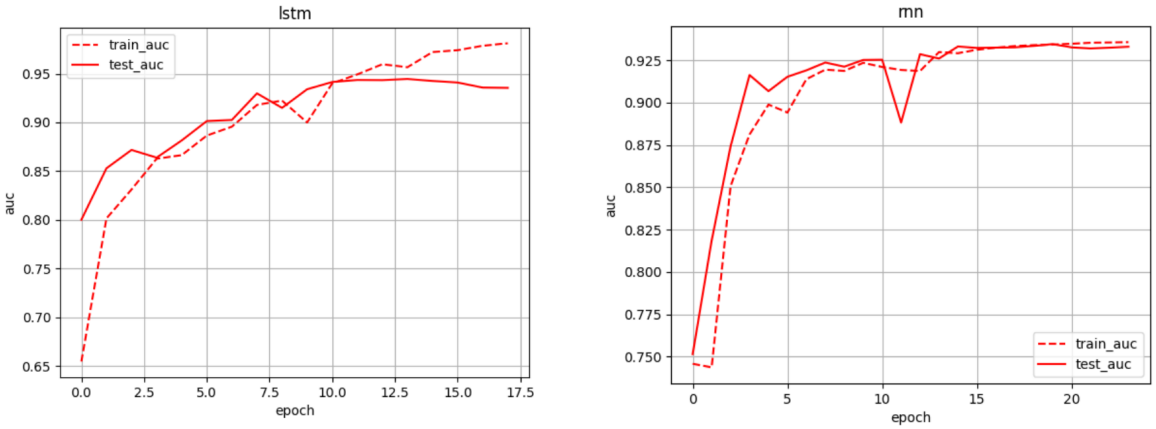


1. 加载经3.3.1处理的语料，将所有语句规范化为统一长度并张量化。由于分词与去除停用词处理后，所有评论以词为单位的最大长度为230，为降低模型计算复杂度，将语句长度设置为200，超过的部分直接截断，不足的部分使用填充。
2. 定义可供模型迭代加载的小批量数据加载器，其中batch_size设为1024。

3. 创建模型并随机初始化权重。如上图所示，模型首先包含1个嵌入层来进行词向量表示，其采用3.3.2节中预训练的Word2vec的第一个嵌入层，该层参数在训练过程中不会发生变化。接着，两个模型的区别主要是RNN模型采用RNN层；而LSTM模型包含LSTM层。最后包含一个全连接层，将结果映射为实数，并使用sigmoid激活函数将结果范围转化为[0,1]。
4. 模型训练。

为了获得更好的结果，本文将循环层数为1、2、3的情况与每层隐藏节点为128、256、512的情况进行组合并进行了多次实验。最终，RNN模型在包含2个RNN层且每层包含128个隐藏节点；LSTM模型在包含3个LSTM层且每层包含512个隐藏节点的情况下表现最佳。实验超参数设置如表所示。

参数名	参数值	参数意义
lr	0.01	初始学习率
embed_size	100	词向量长度
batch_size	1024	每批运行样本数

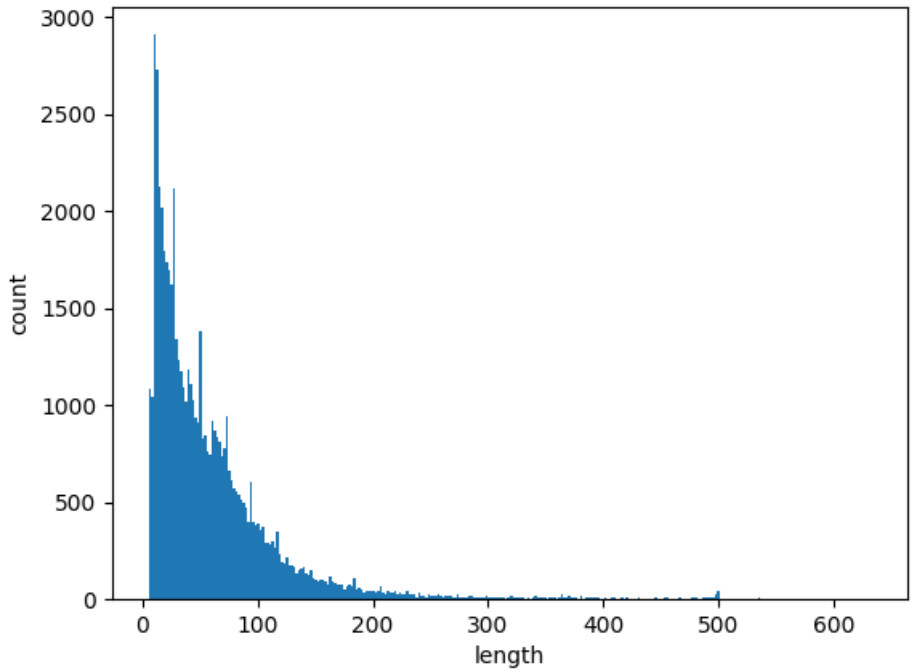


如上图所示，基于RNN的模型迭代了23个epoch，最大AUC出现在第19轮，值为0.9345；基于LSTM的模型迭代了17个epoch，最大AUC出现在第13轮，值为0.9445。lstm模型出现了较为严重的过拟合。

3.4 基于BERT的商品评价分析模型

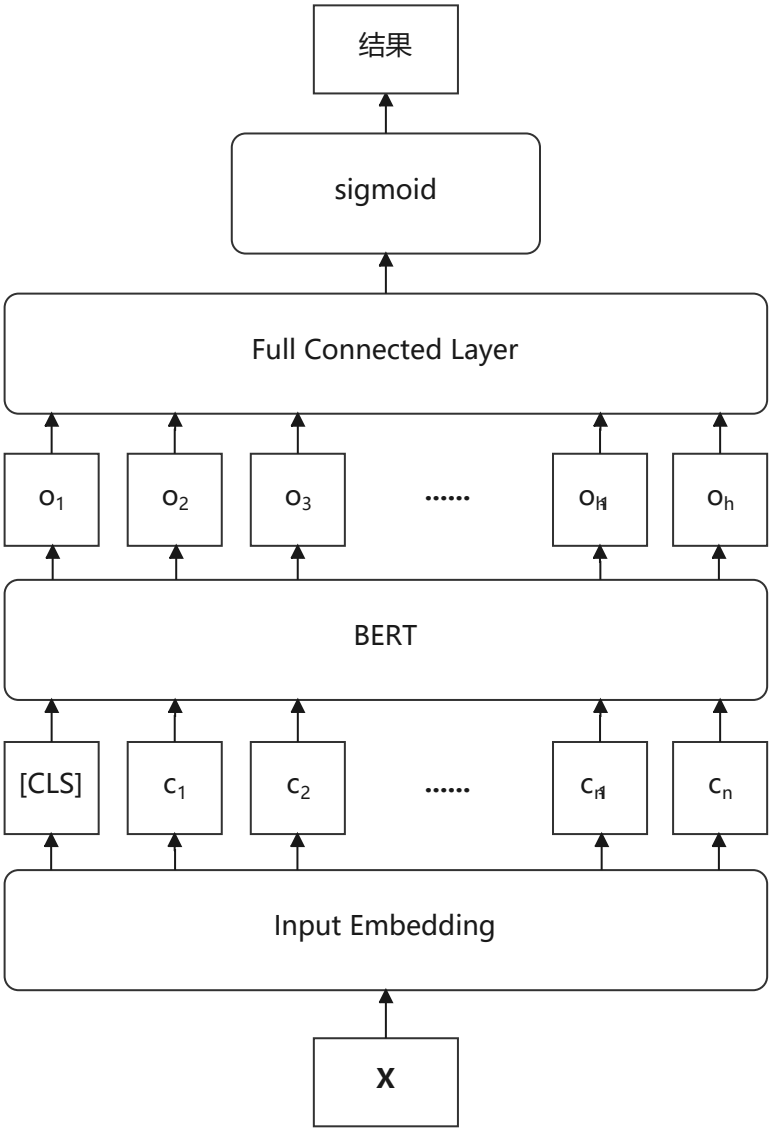
当处理的文本是中文时，BERT以字作为基本单位。本节继续采用经3.2节处理的语料进行模型训练。此外，本文所采用的BERT模型是基于中文维基百科语料预训练过的$BERT_{BASE}$模型，因此仍然使用相应字典，该字典共包含19217个字符。

下图展示了语料中所有评论长度的分布情况，可以看到，绝大部分句子的长度都在300以内，达到总数的98.6%。为了减少模型计算复杂度，在之后的处理中，对于评论长度超过300的部分将直接截断，不足的部分使用特殊符号填充。

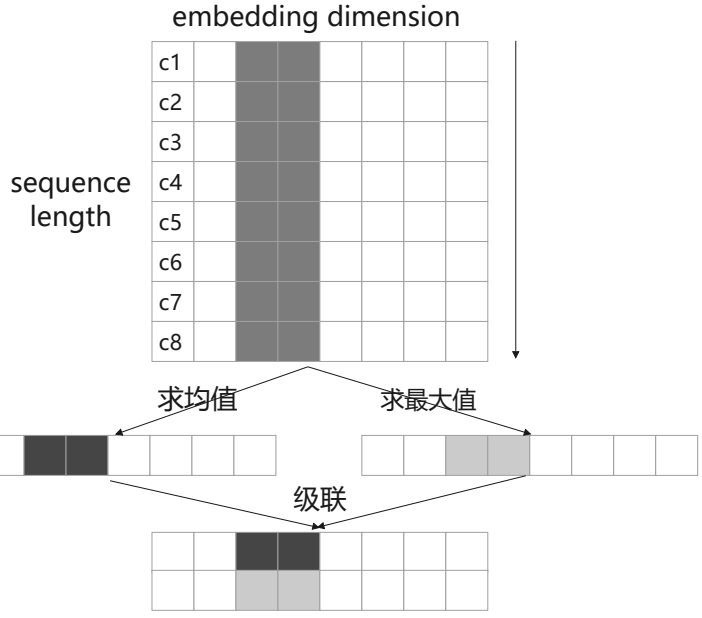


接着，构建模型训练时可迭代加载的Dataset，加载所有语料与标签。为了防止过拟合，将训练集10%的语句以中文分隔符为间隔进行再分句。然后，将所有句子使用字典向量化，添加与，并将超出部分截断。返回token_embeddings以及对应的标签label。

基于BERT的商品评价分析模型结构如图所示。



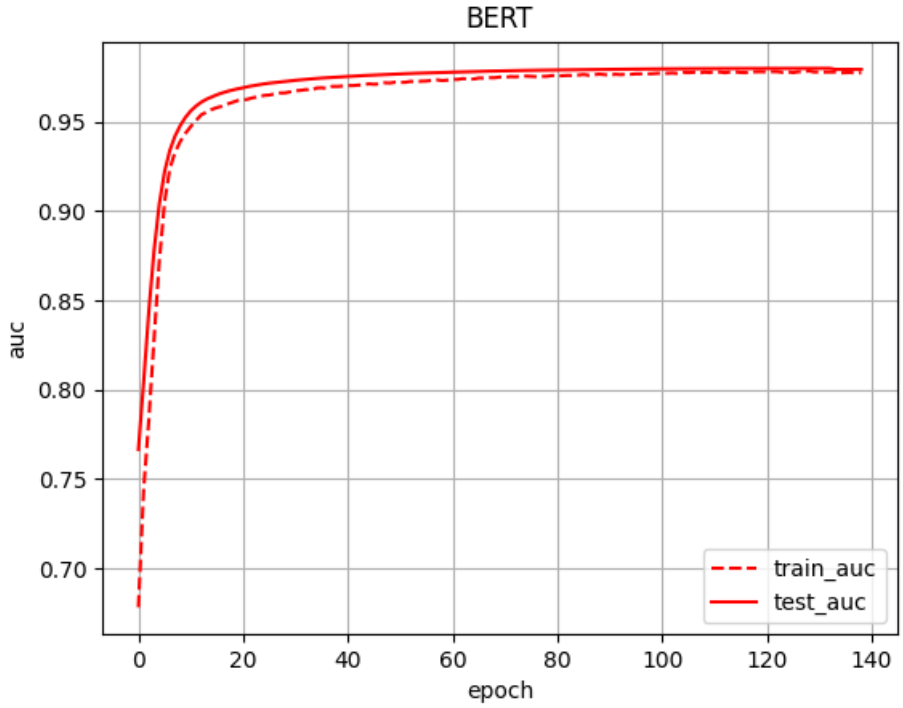
在BERT层之后，附加两个全连接层，对于最后一层的输出采用sigmoid激活函数进行处理，保证输出结果在[0, 1]内。在初始构建时，基于BERT的模型仅添加一个全连接层并使用sigmoid激活函数进行处理，训练结果的AUC值很高，但泛化能力比较差。为了提高模型的泛化能力，使用mean_max_pool进行了一定改进。



max_pooling 和 mean_pooling 就是沿着 sequence_length 这个 维度求最大值和平均值, 之后将这两者级联, 最终结果维度为[batch_size, embed_dim * 2]。 其反映了该句话的平均响应和最大响应。

基于 BERT 的模型的参数设置如表所示。

参数名	参数值	参数意义
lr	1e-06	初始学习率
batch_size	64	每批运行样本数
num_layer	6	Transformer Block 的个数
num_hidden	384	每层隐藏节点数
num_attention_head	12	注意力机制头的个数
dropout	0.4	训练时使参数失活的比例



基于BERT的模型共迭代了138个epoch，测试集最大AUC出现在第133轮，值为0.9797，最佳阈值为0.2。

3.5 模型测试及特殊说明

从3.3与3.4中可以看到，3个模型在训练数据上均达到了非常好的效果，这里采用表现最好的BERT模型进行测试，处理结果如下：

```
----------
骁龙8gen1是真的烫。摄像头是真的突出，不太行
负样本, 输出值0.00
----------
千万别买，伤眼睛不说是真的不好用
负样本, 输出值0.00
----------
手机确实是太好用了，系统特别好用，流畅的不得了
正样本, 输出值0.97
```

3.3节与3.4节中使用了一些方法来优化模型的训练，主要包括如下几方面。

- 动态学习率与早停：采用该方法的主要目的是防止模型由于使用的训练数据相对较少而产生严重的过拟合问题。具体执行步骤如下：
    - 设置耐受阈值patient，动态学习率dynamic_lr以及早停阈值threshold。
    - 在每一轮测试集迭代完成之后，计算本轮的评价指标AUC。若当前AUC小于所有轮的最大AUC，则threshold加1，并调整dynamic_lr，将其变为原来的0.8倍；若当前AUC大于等于最大AUC，则将threshold置0。
    - 如果threshold>patient，则停止训练，否则执行步骤2。
- 寻找最佳分类边界。sigmoid激活函数的输出结果为一个大小在[0,1]的实数，因此需要选择一个合适的阈值将结果划分为两个类别。常见的划分方式是以0.5为基准进行分类，但该种分类方式并不能让模型达到最佳的分类效果。本文以0.01为步长，将范围[0,1]的阈值进行枚举，选择使AUC最大的阈值作为最终的分类阈值。

此外，上述三个模型的训练环境如下表所示。

运行环境	**名称**
操作系统	Linux
显卡参数	RTX 3090
编程语言	Python
深度学习框架	PyTorch