吉林大学珠海学院

**毕 业 论 文**

**基于LCF-ATEPC模型的文本情感分类算法的研究与实现**

学 院： 阿里云大数据应用学院

专 业 名 称： 数据科学与大数据技术

学 生 姓 名： 郑浩

学 号： 20190305

指导教师姓名、职称： 罗鸿雁 教授

完成日期：2023年 5月 14日

基于LCF-ATEPC模型的文本情感分类算法的研究与实现

摘 要

随着社交网络的普及，越来越多的人在网上表达自己的想法，导致全球信息量呈现出爆炸式的增长趋势。把握这些海量信息的情感信息对于企业、政府以及社会的决策和发展具有重要意义。因此，文本情感分类技术得到了越来越多的关注。目前，基于深度学习的文本情感分类算法已经成为主流，并在不断地发展和优化，这种技术在商品评价、舆情分析、情感监控等领域得到了广泛的应用，拥有重要的研究和实际应用意义。

本文针对文本情感分类问题，使用了LCF-ATEPC模型，是一种面向中文及多语言的ATE(Aspect Term Extraction,属性提取)和APC(Aspect Polarity Classification,属性情感分类)联合学习模型，可用于解决分析文本评论中特定方面情感的任务，即识别评论中的方面词汇并将其分类为正面、负面或中性情感。该模型以BERT模型为基础，通过采用多层上下文融合机制、方面词汇注意力机制和多任务学习技术，进一步提高了方面级情感分析的准确性和效果。

传统的情感分析，一般是针对整个句子的，比如购物网站上的好评或者差评。但在实际场景中，对于某个商品的评价，往往是有好有坏的。所以我们使用了LCF-ATEPC模型，实现了对多个实体的评论文本的情感分类，即将好的和坏的评价对象区分评判。

关键词：情感分析；多头注意力机制；LCF-ATEPC；ATE；APC；CRF；预训练BERT模型；CDM；CDW；

Research and Implementation of Text Emotion Classification Algorithm Based on LCF-ATEPC Model

Abstract

With the popularization of social networks, more and more people express their ideas online, leading to an explosive growth trend in global information volume. Grasping the emotional information of these massive amounts of information is of great significance for the decision-making and development of enterprises, governments, and society. Therefore, text sentiment classification technology has received increasing attention. At present, deep learning based text sentiment classification algorithms have become mainstream and are constantly developing and optimizing. This technology has been widely applied in fields such as product evaluation, public opinion analysis, sentiment monitoring, and has important research and practical application significance.

This article focuses on the issue of text sentiment classification and uses the LCF-ATEPC model, which is a joint learning model for Chinese and multilingual ATE (Aspect Term Extraction) and APC (Aspect Polarity Classification). It can be used to solve the task of analyzing specific aspect emotions in text comments, that is, identifying aspect vocabulary in comments and classifying them as positive, negative, or neutral emotions. This model is based on the BERT model and further improves the accuracy and effectiveness of aspect level sentiment analysis by adopting a multi-layer context fusion mechanism, aspect vocabulary attention mechanism, and multi task learning technology.

Traditional emotional analysis typically focuses on the entire sentence, such as positive or negative reviews on shopping websites. However, in practical scenarios, the evaluation of a certain product is often mixed. So we used the LCF-ATEPC model to achieve emotional classification of comment texts from multiple entities, which is to distinguish and evaluate good and bad evaluation objects.

Key words：Emotional analysis; Multi-head self-attention; LCF-ATEPC; ATE; APC; BERT-Shared; CDM; CDW;

目 录

[1 绪 论 1](#_Toc133096053)

[1.1 课题研究的背景及意义 1](#_Toc133096054)

[1.2国内外研究现状 1](#_Toc133096055)

[1.3课题的主要研究内容 2](#_Toc133096056)

[1.4论文章节安排 2](#_Toc133096057)

[2 基于LCF-ATERPC模型的文本情感分类算法关键技术研究 3](#_Toc133096058)

[2.1 LCF-ATEPC模型 3](#_Toc133096059)

[2.1.1方法架构 3](#_Toc133096060)

[2.1.2 BERT层 4](#_Toc133096061)

[2.1.3 语义相对距离（SRD） 4](#_Toc133096062)

[2.1.4上下文特征动态掩码（CDM） 5](#_Toc133096063)

[2.1.5上下文特征动态加权（CDW） 5](#_Toc133096064)

[2.1.6多头注意力机制（MHSA） 6](#_Toc133096065)

[2.1.7 Aspect项提取 6](#_Toc133096066)

[2.1.8 BIO标签 7](#_Toc133096067)

[2.1.9 ATE/APC 7](#_Toc133096068)

[2.2 本章小结 7](#_Toc133096069)

[3 短文本情感分析 8](#_Toc133096070)

[3.1模型架构 8](#_Toc133096071)

[3.2数据预处理 8](#_Toc133096072)

[3.2.1 合并样本数据 9](#_Toc133096073)

[3.2.2 异常值剔除 10](#_Toc133096074)

[3.3 Bert Tokenizer分词和编码 10](#_Toc133096075)

[3.4 BIO标签和情感值的转化 11](#_Toc133096076)

[3.5 CDM/CDW 11](#_Toc133096077)

[3.5.1 解析实体位置 11](#_Toc133096078)

[3.5.2 计算实体权重参数 11](#_Toc133096079)

[3.5.3 数据处理 12](#_Toc133096080)

[3.6 CRF层 13](#_Toc133096081)

[3.6.1 Bert\_CRF 实体模型初始化 13](#_Toc133096082)

[3.6.2实体标签预测 13](#_Toc133096083)

[3.7 Self-Attention 14](#_Toc133096084)

[3.8解析句子中多个实体的情感分类 14](#_Toc133096085)

[3.8.1 损失值计算 14](#_Toc133096086)

[3.8.2 模型训练 15](#_Toc133096087)

[3.8.3 情感预测 16](#_Toc133096088)

[3.9本章小结 17](#_Toc133096089)

[4 实验及结果分析 18](#_Toc133096090)

[4.1评估指标 18](#_Toc133096091)

[4.1.1评估指标介绍 18](#_Toc133096092)

[4.1.2实验结果 18](#_Toc133096093)

[4.2模型测试 19](#_Toc133096094)

[4.3本章小结 19](#_Toc133096095)

[5 总结与展望 20](#_Toc133096096)

[5.1总结 20](#_Toc133096097)

[5.2展望 20](#_Toc133096098)

[参考文献 21](#_Toc133096099)

[致 谢 22](#_Toc133096100)

# 

# 1 绪 论

## 1.1 课题研究的背景及意义

随着互联网的进一步普及，我们生活中产生的用户生成文本（UGC）越来越多，包括评论、微博、博客等等。这些文本记录了我们对世界的观察、想法和情感。这种UGC的大量产生，使得我们有了更多的表达和沟通的渠道，同时也为企业和政府等组织提供了更多的信息来源。然而，随着UGC的不断增长，人工处理和分析这些文本已经变得越来越困难。这就需要计算机自动化处理这些文本，把它们转化为有用的信息。情感分析技术应运而生，成为处理这些文本的重要手段。

文本情感分析[4]，是指利用自然语言处理技术和机器学习算法等方法，对文本数据中的情感信息进行识别和提取的过程。是一项重要的自然语言处理（NLP）技术，它可以帮助人们更好地理解和分析文本中的情感信息。文本情感分析可以应用于各个领域，例如社交媒体、品牌管理、舆情监测、客户服务等。在社交媒体方面，文本情感分析可以帮助企业了解用户对其品牌或产品的态度和看法，以及用户的需求和期望。在品牌管理方面，文本情感分析可以帮助企业评估其品牌声誉和形象，从而更好地制定品牌策略。在舆情监测方面，文本情感分析可以帮助政府和企业了解公众对某个事件或话题的态度和看法，从而更好地制定相关政策和决策。在客户服务方面，文本情感分析可以帮助企业快速识别和回应用户的需求和问题，从而提高客户满意度和忠诚度。总之，文本情感分析是一项有广泛应用前景的技术，将在未来的各个领域得到更广泛和深入的应用。因此，对这些短文本进行情感分析具有很大的研究意义和实用价值。

## 1.2国内外研究现状

近年来，人工智能技术飞速发展，其中情感分析技术也不断涌现出新的方法和应用。准确分析文本数据对于情感分析的研究至关重要，只有准确理解文本所表达的情感倾向，计算机才能够正确地学习和实现情感分类。因此，自情感分析技术提出以来，越来越多的学者开始关注和研究这个领域。目前，情感分析技术主要有三种，即基于情感知识、机器学习和深度学习。这些方法的不断发展和创新，使得情感分析技术在社会和企业的决策、用户体验、舆情监控等方面得到了广泛的应用和推广。

## 1.3课题的主要研究内容

在当今社会，随着社交媒体、电商平台等互联网应用的普及，大量的用户产生了短文本，如微博、评论、短信等。这些文本的特点是长度短，且一句话中包含多种情感，给文本情感分析带来了巨大挑战。

为了从这些数据中得到有价值的信息，本文使用了LCF-ATEPC模型[1]，是一种基于注意力机制的远程监督方案，用于解决中文情感分析和命名实体识别（NER）的问题。该模型结合了词级别和句子级别的信息，并使用了ATE和APC两个子任务来进行情感分析。

具体来说，该模型首先使用预训练BERT模型对文本进行编码，得到文本的表示。然后，使用ATE子任务来识别文本中的特定方面术语，并通过句子级别的注意力机制来捕捉方面术语和文本其他部分之间的关系。接下来，使用APC子任务来确定每个方面术语所属的情感类别，同时使用词级别的注意力机制来区分不同方面术语之间的情感极性。最后，将ATE和APC两个子任务的结果组合起来，即可得到对整个文本的情感分析结果。

该模型的使用，可以帮助企业了解消费者的满意度和不满意的原因。提供方面级别情感分析：将用户评论中提到的方面（如价格、质量、服务等）与相应情感极性进行匹配，帮助企业了解消费者对不同方面的满意度和不满意的原因。提供实体级别情感分析：在方面级别情感分析的基础上，提取出评论中涉及的具体实体（如产品型号、商家名称等）并进行情感分析，帮助企业了解消费者对不同实体的评价和需求。进而优化产品设计、提升服务质量、改进营销策略等，从而提升品牌形象和市场竞争力。

## 1.4论文章节安排

本论文的主要章节安排如下：

第一章：绪论。首先探讨了互联网的发展趋势，然后介绍了情感分析技术的背景及其在实际应用中的重要性。

第二章： 情感分析的相关理论介绍。本章首先介绍了LCF-ATEPC的模型架构，其次介绍了模型中各层的原理和各项参数的计算方法。

第三章： 实验的实施细节的介绍，阐述了模型运行的流程，以及数据经过各层的结果。

第四章：介绍评估模型效率的各项指标和训练结束后模型的最终表现。

第五章：总结与展望，对本次实验的研究内容进行了总结，同时也发现了一些需要进一步思考和优化的问题，对未来研究的发展方向进行了预测和展望。

# 2 基于LCF-ATERPC模型的文本情感分类算法关键技术研究

情感分析是NLP领域的主要研究任务之一，在丰富的理论基础支撑下，已经取得了很多卓越的研究成果。在本章节中，我们对情感分析任务的相关理论知识进行了阐述，并对该领域的相关工作进行了分析和总结，从而为后续研究的工作做好铺垫。

## 2.1 LCF-ATEPC模型

该模型是一种基于情感分类的联合任务模型，采用了BERT-ADA领域的自适应BERT模型（旨在解决预训练模型在不同任务上的适应性问题）并将局部上下文焦点机制[5]整合到模型中。本节介绍LCF-ATEPC模型的架构和方法。

### 2.1.1方法架构

该模型将ATE和APC任务相结合，利用两个独立的BERT层分别对全局和局部上下文进行建模，并采用多任务学习的方法，将输入序列标记为不同的标签，以同时完成两个任务。第一种标签指示令牌是否是方面词，第二种标签标记方面词的情感极性。

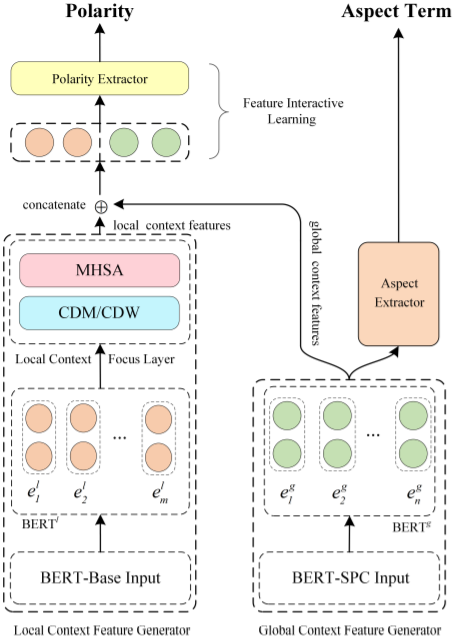


图2.1 LCF-ATEPC模型架构

由图2.1所示，位于右侧的单元是GCFG（Global Context Feature Generator，全局上下文特征生成器），包含一个独立的预训练BERT层，目的是为了获取分词ID。表示经过Bert编码得到768维词向量。GCFG 单元仅部署一个 MHSA 编码器来获取全局上下文特征。Aspect extractor（方面提取）[3]表示通过BIO标记获取BIO标签。

位于左侧的单元是LCFG（Local Context Feature Generator，局部上下文特征生成器），包含一个独立的预训练BERT层，目的是为了获取分词ID。表示经过Bert编码得到768维词向量。CDM（context-feature dynamic mask，情感检测）用于检测句子中情感表达，CDW（context-feature dynamic weighting，动态权重）融合上下文和特征信息来预测文本标签，这一层目的是为了分别得到特征权重。MHSA（multi-head self-attention，多头注意力机制），将上一层得到的两个信息进行混合，得到一个新的特征矩阵。FIL（Feature Interactive Learning，特征交互学习）将MHSA得到的某个实体特征和全局特征进行拼接，最后进行情感分类。

### 2.1.2 BERT层

预训练的BERT模型[10]旨在提高NLP任务的性能，LCF-ATEPC模型使用了两个独立的BERT-Shared层，以提取实体和全局上下文的特征。

(2-1)

(2-2)

公式(2-1)和(2-2)中的和分别是 LCFG 和 GCFG 的输出特征。和分别是嵌入在LCFG和GCFG中的相应BERT共享层。和分别表示 LCFG 和 GCFG 的标记化输入，我们可以得到局部和全局上下文特征的初步输出。

### 2.1.3 语义相对距离（SRD）

局部上下文的确定依赖于语义相对距离，SRD用来判断上下文是否属于目标方面的局部上下文。在现有的ABSA（文本分析）中，通常将输入序列划分为方面序列和上下文序列，并将方面词和上下文视为不同的片段，分别对它们进行建模，但是该模型与以往不同，没有将aspect（方面）作为单独的输入，而是挖掘方面及其局部上下文。

SRD是一个基于token-aspect对的概念，描述token与aspect之间的距离，就是token与aspect之间隔了几个token。

(2-3)

公式(2-3)中i是特定令牌的位置，是aspect的中心位置，m 是目标aspect的长度，所以SRD表示第i个token和目标aspect之间的距离。

### 2.1.4上下文特征动态掩码（CDM）

CDM的核心思想是利用上下文信息来增强方面词的表示[6]，从而提高方面级情感分析的性能。除了局部上下文特征，CDM将屏蔽层学习到的非局部上下文特征，即将非局部上下文的所有位置的特征设置为零向量。

假设是的初始输出特征，那我们可以得到如下的局部上下文特征。

(2-4)

公式(2-4)中M是用于屏蔽非局部上下文特征的屏蔽矩阵，是输入序列中每个token的掩码向量。α是SRD阈值，n是包括方面的输入序列的长度。其中与目标aspect相关的SRD小于阈值 α表示的是局部上下文。E是一个1向量，O是一个零向量（长度是n）。

将学习到的局部特征放到一个MHSA层（求局部上下文中的注意力）中，就得到了局部上下文表示 。

(2-5)

### 2.1.5上下文特征动态加权（CDW）

CDW的核心思想是通过在上下文中考虑单词的语义特征来增强情感分析和情感推理的准确性。与CDM层相比，CDW采用了一种更加温和的策略，对目标aspect的非局部上下文特征根据SRD进行加权衰减。局部上下文特征保持不变。

假设是的初始输出特征，那我们可以得到如下的局部上下文特征。

(2-6)

公式(2-6)中W是权重矩阵，表示的是权重向量。α是SRD阈值，n是包括方面的输入序列的长度。其中与目标aspect相关的SRD小于阈值 α表示的是局部上下文。E是一个1向量，O是一个零向量（长度是n）。

将学习到的局部特征放到一个MHSA层中，就得到了局部上下文表示。

(2-7)

CDM和CDW层是独立的，这就意味着他们是可选的，他们的输出特征均是。因此该模型将CDM和CDW层的学习特征串联起来，并将他们的线性变换作为局部上下文的特征。

(2-8)

公式(2-8)中, *,* 表示的是权重矩阵，表示的是偏移矩阵。

### 2.1.6多头注意力机制（MHSA）

MHSA是多个self-attention同时工作，将获得的结果进行处理，用来提取语境中的深层语义特征,最后获得的是注意力分数矩阵[7]。MHSA可以有效地避免上下文之间的长距离依赖关系所带来的负面影响，在学习特征时具有优异的性能。

对于每一个向量x，分别乘上三个系数，，，得到q,k,v三个值：

写成向量形式：

写成向量形式： (2-9)

写成向量形式：

得到的Q,K,V分别表示querys(要去查找的),keys(等待被查找的)和values(真实的特征信息)。计算q和k的内积可以得到它们之间的匹配程度，输入两个向量得到一个分值score，最终的得分值经过softmax即()，就是最终上下文结果。

### 2.1.7 Aspect项提取

是方面级情感分析[2]的重要组成部分，方面项提取器首先对每个token（令牌）执行token-level分类，假设 是令牌T对应位置上的特征（通过一个线性层分类器，转换成）。每个位置都相当于一个三分类问题。

(2-10)

公式(2-10)中,N是令牌类别的数量，表示方面极性分类器推断的令牌类别。

### 2.1.8 BIO标签

BIO标签是指将命名实体识别问题（Named Entity Recognition，NER）中的实体边界信息表示出来的一种方式。在BIO标注中，B代表实体的开始（Beginning），I代表实体的中间（Inside），O代表实体外（Outside）。

在一个句子中，如果一个词是实体的开头，那么就用“B-实体类型”的标签来标记这个词，如“B-PER”代表人名实体的开头。如果一个词是实体的中间部分，那么就用“I-实体类型”的标签来标记，如“I-PER”代表人名实体的中间部分。如果一个词不是实体，那么就用“O”来标记。

### 2.1.9 ATE/APC

ATE（Aspect Term Extraction）[8]指的是从文本中提取方面词的相关知识。在该模型中，ATE模块是用于提取句子中与方面词相关的单词，以及确定方面词的位置。这个模块使用预先训练好的BERT模型来提取方面词，输出是一个方面词序列，该序列标识了文本中所有方面词的位置。

APC（Aspect Polarity Classification）[9]模块的主要任务是对从ATE模块中提取出来的方面词（aspect term）进行情感极性分类，即判断方面词所表示的情感是正向的、负向的还是中性的。

## 2.2 本章小结

在这一章中，我们介绍了本文所涉及的情感分析模型：LCF-ATEPC，详细介绍模型架构、Bert层、CDW/CDM层、多头注意力机制、方面提取、BIO标签的定义、计算步骤和计算公式，其计算方法将直接用于后续实验。

# 3 短文本情感分析

在本章节中，主要阐述关于情感分析任务的具体操作流程和各层间的核心代码解释。

## 3.1模型架构

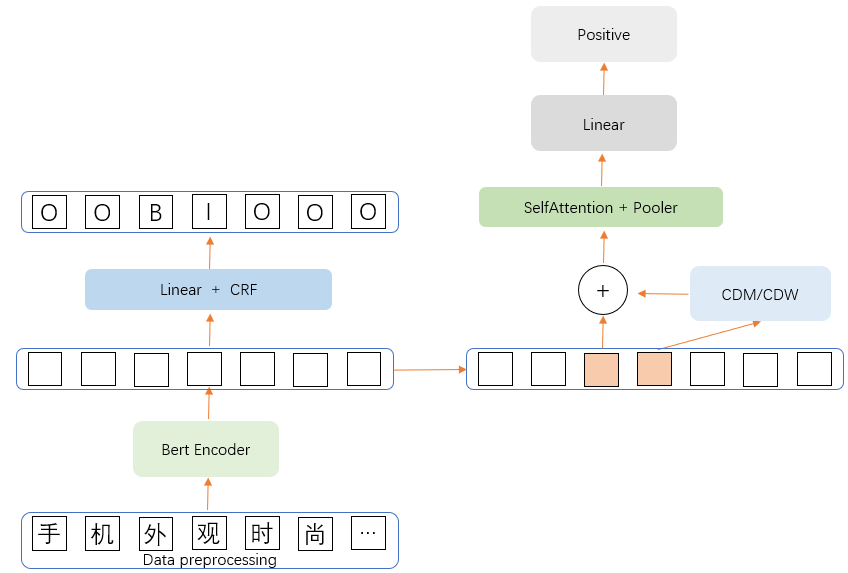


图3.1 LCF-ATEPC模型流程图

如图3.1，以“手机外观时尚”这句话为例，数据预处理以后每个字以空格隔开，经过Bert分词和编码得到768维词向量，并且复制成两个部分。左侧部分进行实体识别，即BIO标记，解析出实体“外观”。右侧部分，先计算CDW，CDM权重并进行叠加，由于情感分类是基于整个句子的，所以要在SelfAttention层进行混合，最后经过Linear层做情感分类。

## 3.2数据预处理

首先利用python爬取京东，淘宝商品评论数据10万条并存为csv格式，并且只保留数据集中的评论序号，评论文本信息。文本信息和分值是情感分析任务中至关重要的两种信息，它们分别对应着训练数据和类别标签。然后经过“标注精灵助手”软件对数据进行标注，一个字对应一个BIO标记和一个分类标签，结果如下图3.2：



图3.2 原始数据

可以看到结果数据一字一行，非常不便于处理，且数据中存在标注错误的情况。因此，还需要完成两个任务，一个是将样本文件处理成一句一行的形式，方便读取，另一个是，剔除异常数据。

### 3.2.1 合并样本数据

处理策略是按空行切分后，合并句子，每行一句，再把不同多个样本文件合并成一个，方便读取。核心代码如下图3.3：

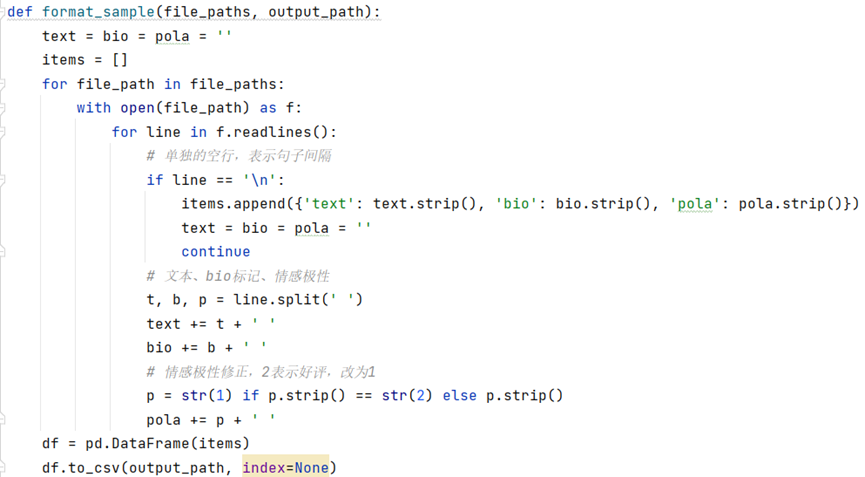


图3.3合并样本数据代码

经过如上方法的处理即可实现一句一行，如下图3.4：

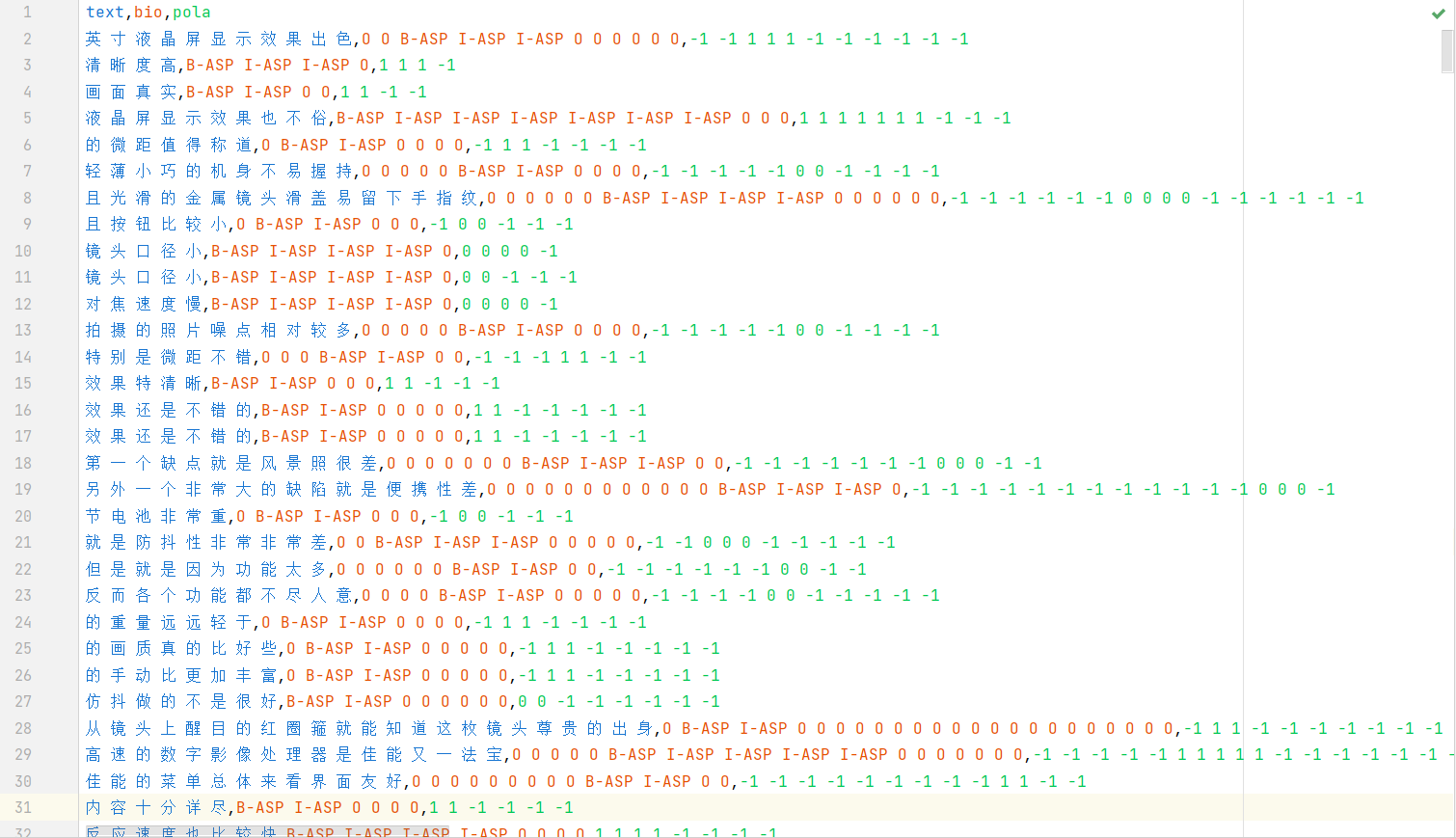


图3.4数据合并结果

### 3.2.2 异常值剔除

正常来说，标记为实体(B-ASP)的文字，对应情感数值，要么好评要么差评，即数值应该是1或2，不应该为-1等其他数字，如图3.5：



图3.5异常数据

处理方法也很简单，即判断B-ASP对应的值即可，如果不为1或2，就删除本条数据。运行结束后，如图3.6可以看到只有四条数据是异常的：

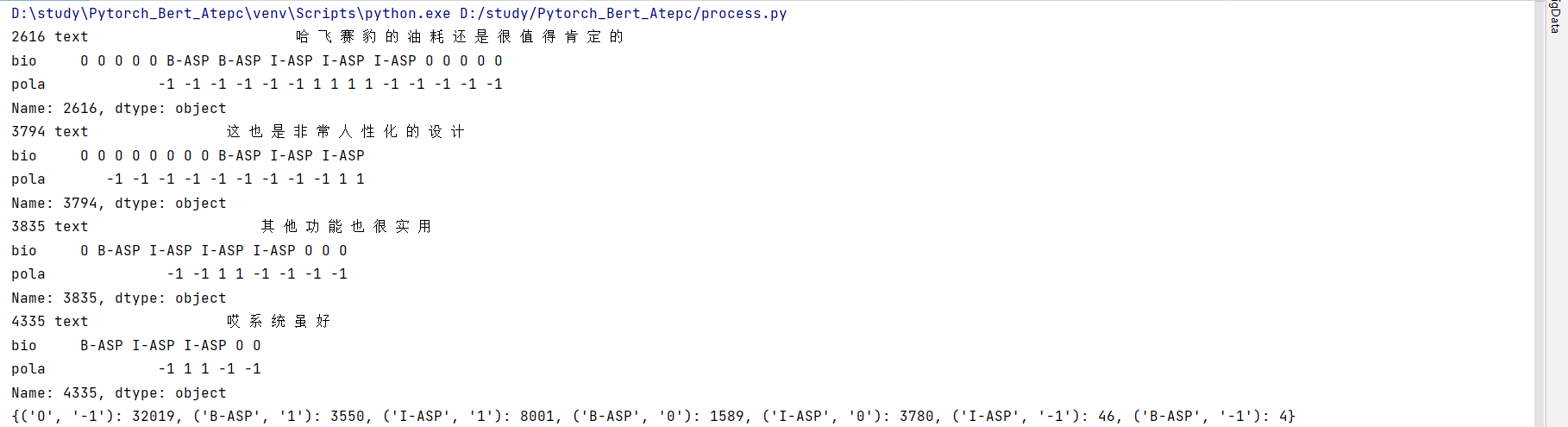


图3.6异常数据数量

## 3.3 Bert Tokenizer分词和编码

在数据清洗之后，需要对数据集中的评论信息进行BERT词向量的提取。使用 BERT 这个预训练模型，主要分为两步：第一步，使用 Bert Tokenizer 进行分词，并将词转化为id；第二步，使用 Bert Model 进行词编码，将每个词编码成 768 维的向量。



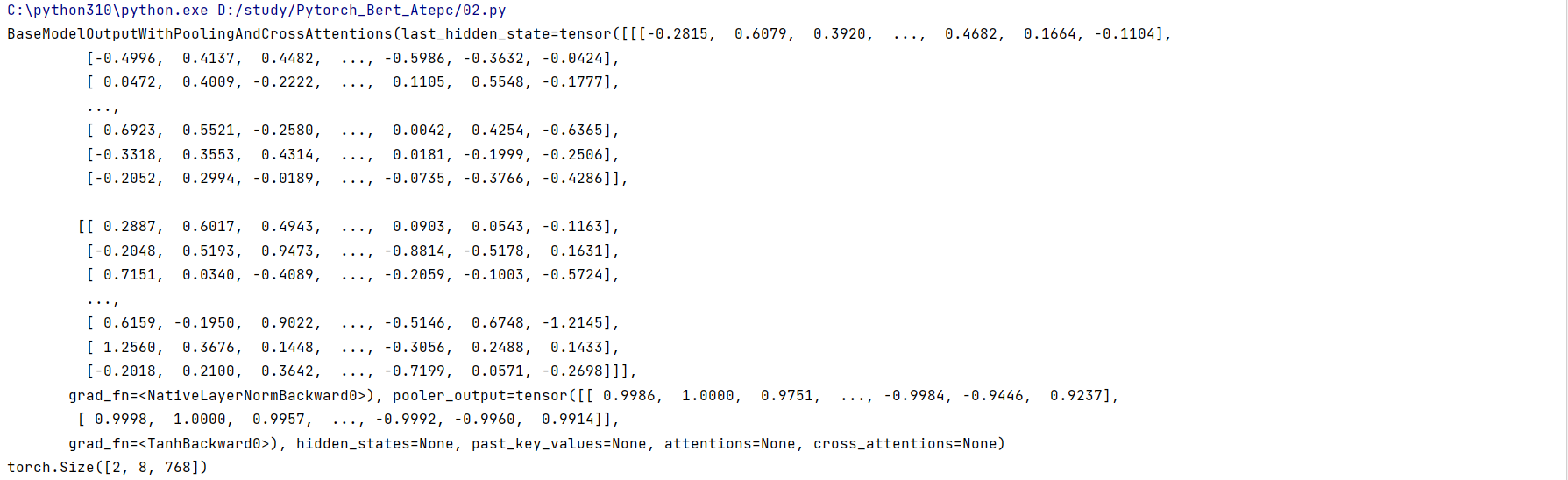


图3.7词向量

如图3.7，以“手机外观时尚”这句话为例，经过Bert Tokenizer分词得到“[CLS][手][机][外][观][时][尚][SEP]”（CLS和SEP分别表示句子的开始和结束），Bert预训练模型可以得到input\_id（词ID），token\_type\_id（句子在数据集中的位置），attention\_mask（是否有填充，不等长填充0）。输入Bert预训练模型，得到last\_hidden\_state 768维词向量。

## 3.4 BIO标签和情感值的转化

接下来需要将BIO标签转化为id（“O”：0，“B-ASP”：1，“I-ASP”：2），情感值转化为数字。



图3.8 BIO标签

如图3.8，以数据集中的第一句“英寸液晶屏显示效果出色，清晰度高”为例，通过第一行的id可以看出该句有两个实体，并且情绪为1（积极）。核心代码如下：

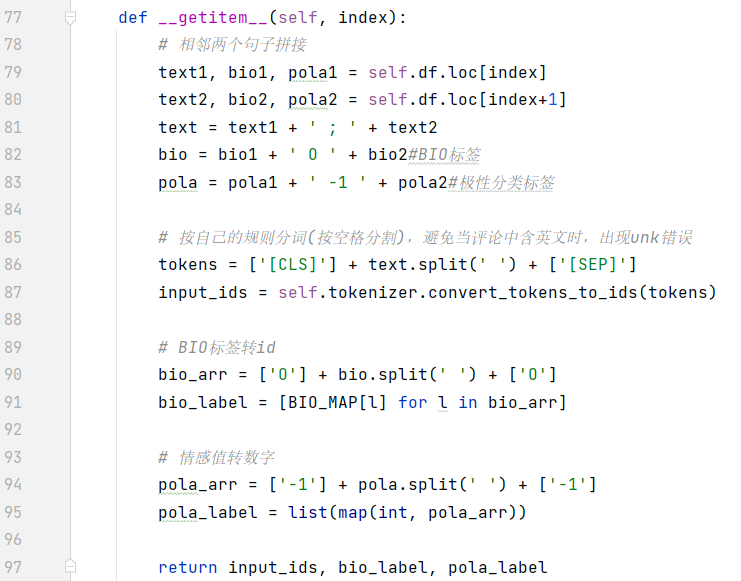


图3.9 BIO标签转化核心代码

## 3.5 CDM/CDW

### 3.5.1 解析实体位置

完成了数据加载和转id的过程，还需要计算CDW和CDM两个权重参数。另外，要实现数据批量加载，还需要将训练样本填充到等长。首先需要找到实体位置，策略是找到编码不为0的位置，例如“[CLS]这个手机外观时尚，美中不足的是拍照像素低。[SEP]”这句话的编码是“0 0 0 0 0 1 2 0 0 0 0 0 0 0 0 0 1 2 2 2 0 0 0”找到位置结果是：



图3.10实体位置

### 3.5.2 计算实体权重参数

实体+SRD距离内的权重为1，权重值和距离成反比，距离越远权重越低。具体思路是，当前位置减去实体的开头和结尾的绝对值。以第一句为例如，我们可以看到离实体越远，权重越小。

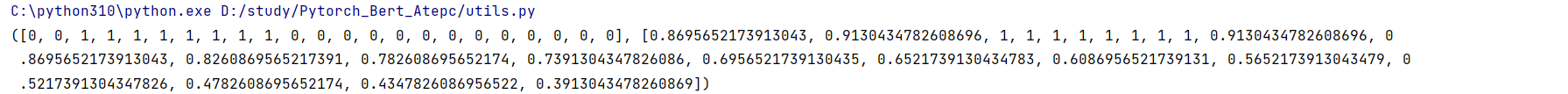


图3.11实体权重

核心代码如下图3.12：



图3.12权重计算核心代码

### 3.5.3 数据处理

首先需要定义最大句子长度，然后对变量进行初始化，需要初始化的参数是input\_id、BIO标签、填充后的mask值、权重参数CDM和CDW、表示实体情况的参数pairs。接下来就需要逐条处理数据，当在处理一个句子时，可能会有多个实体，为了防止模型计算错误，需要随机取一个，来判断情感极性，另一个也可能在其他轮次中取到。先要获取实体位置，没有实体则跳过，填充句子长度，使得三个参数等长，接下来对实体和情感进行分类并计算权重参数，最后计算情感分类。

以“英寸液晶屏显示效果出色，清晰度高”为例，如图3.13可以看到情感分类为1(积极)。

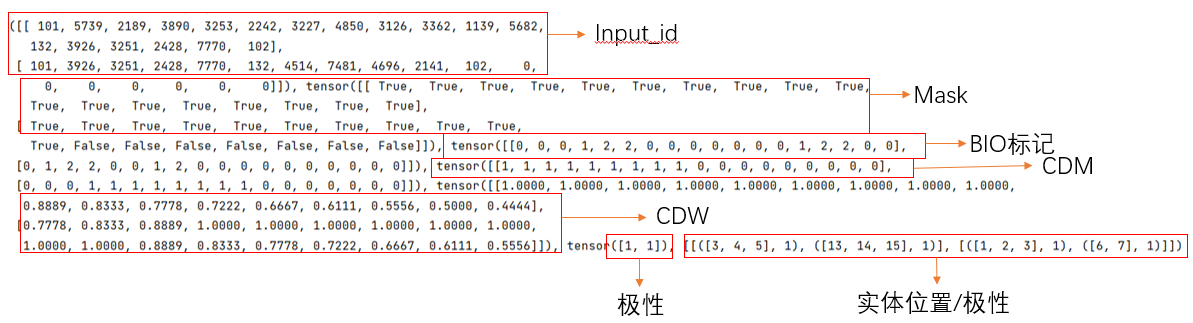


图3.13极性分类

## 3.6 CRF层

### 3.6.1 Bert\_CRF 实体模型初始化

添加完配置项和导入模块后，要想运行Bert实体首先要对模型进行初始化。具体思路是，使用BertModel把模型加载进来，然后冻结Bert参数，只训练下游模型，然后self.ent\_linear把768维向量映射到BIO的三维向量中，最后用CRF矫正BIO标签。核心代码如下图3.14：



图3.14模型初始化

### 3.6.2实体标签预测

为了加强准确率，需要在原有模型中增加一层CRF。首先进入get\_entity方法向get\_text\_encoded方法传入self，input\_id，mask经过Bert编码获取每一个字的词向量，然后把词向量传到linear中，映射到三个向量中后，CRF矫正得到标签的预测值。

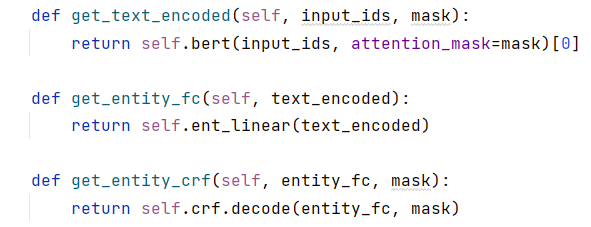


图3.15 实体标签预测核心代码

以“torch.randint(0, 3000, (2, 30))”（生成随机整数范围是0-3000，每个batch两个句子，每个句子30个字）为例预测到的实体标签是：

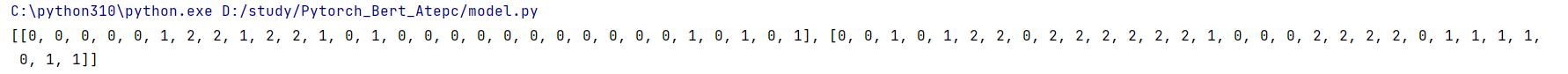


图3.16实体标签预测

## 3.7 Self-Attention

先把权重参数扩维到和词向量一样的维度，再相乘，得到带权重信息的词向量。在CDM进行升维然后横向复制，得到b\*c\*768的向量，CDW同理，然后将CDM和CDW两个向量相乘得到feature新的特征矩阵。然后在768位置进行拼接，即text\_encoded, cdm\_feature, cdw\_feature得到768x3的维度，然后进过linear压到768维。然后self-attension 结合上下文信息，增强语义，pooler 取[CLS]标记位，作为整个句子的特征。最后得到一个二维向量然后用sigmoid进行激活。核心代码如下图3.17：

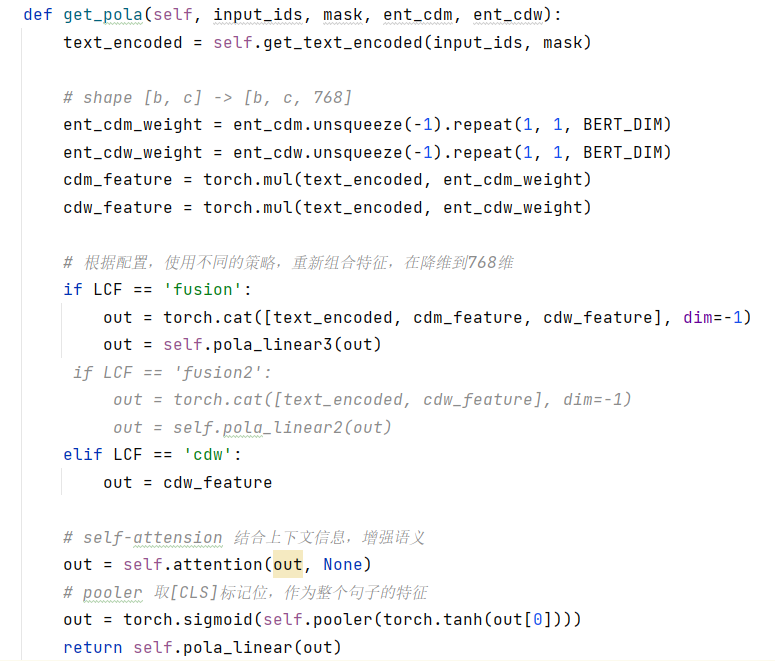


图3.17情感标签预测核心代码

以上文中预测的实体标签和随机生成的CDW,CDM（“ent\_cdm = torch.rand((2, 30)”,“ent\_cdw = torch.rand((2, 30)”）为例，可以得到如图3.18二维向量（情感值）：



图3.18预测结果

## 3.8解析句子中多个实体的情感分类

### 3.8.1 损失值计算

损失值的计算是两部分的和，一部分是实体识别部分的损失，另一部分是情感预测部分的损失。

首先需要传入input\_ids, ent\_label, mask, pred\_pola（预测分类）, pola\_label（真实分类）这几个参数后，分两部分计算损失。第一部分是ent\_loss（实体损失）计算，首先输入参数后经过BERT，然后通过linear转化，然后传到CRF中去计算损失(返回损失是一个平均值)。第二部分是pola\_loss（情感预测损失），即用 pred\_pola, pola\_label做交叉熵损失。核心代码如下图3.19：

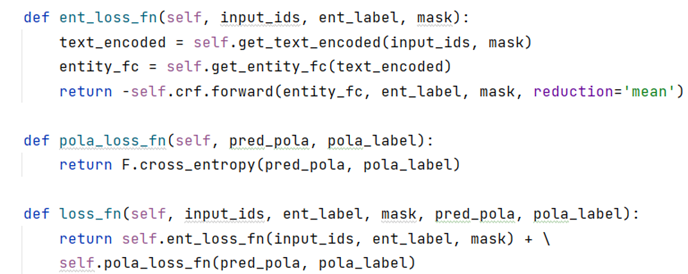


图3.19损失计算核心代码

以第一个batch为例，如图3.20得到的损失值为



图3.20损失计算结果

### 3.8.2 模型训练

我们首先Model实例化，用Adam（梯度下降算法）定义一个优化器，Dataset数据集实例化，使用loader加载，然后分别训练input\_ids, mask, ent\_label, ent\_cdm, ent\_cdw, pola\_label，然后实体预测，情感极性预测，计算损失，梯度清零，反向传播，计算当前梯度，根据梯度更新网络参数。核心代码如下图3.21：

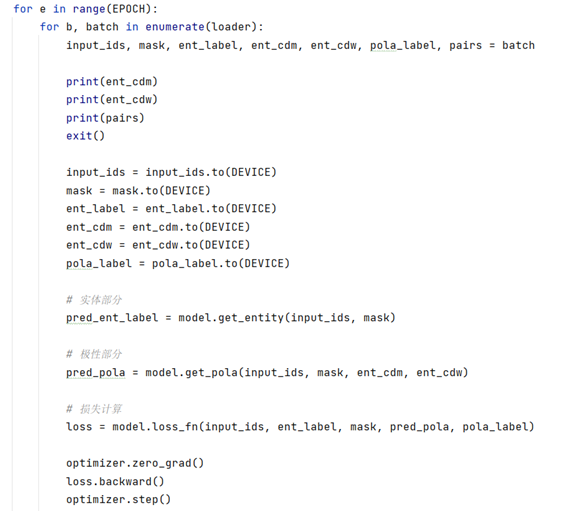


图3.21模型训练核心代码

如图3.22我们可以看到损失值在逐渐下降。



图3.22模型训练结果

### 3.8.3 情感预测

完成了实体标签预测，和某一个实体对应的情感预测，并且计算了联合损失。接下来，需要根据实体的预测值，解析出对应的实体位置，并预测实体对应的情感分类，再跟真实实体对进行对比。

首先要变量初始化，当处理句子时可能会出现句子只有一句，但是实体可能有多个的情况，所以要以实体数量为维度批量处理，根据label解析实体位置，然后n个实体一起预测（同一个句子复制n份，作为一个batch），把列表转换成tensor，为了更进一步得到实体位置的索引值，使用argmax函数（横向取二维数据的最大值），zip函数，使实体位置和分类呈现为一对一的格式。以数据集中的第一句为例，输入如下图3.23参数：

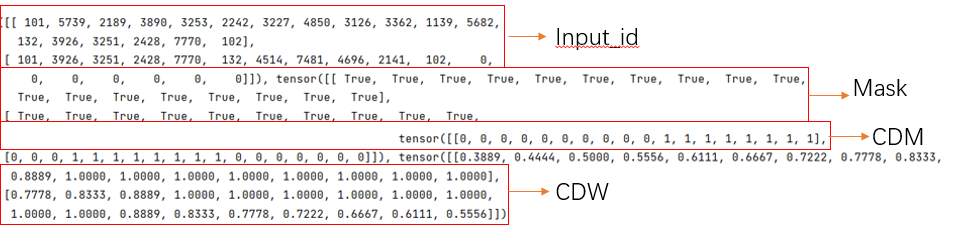


图3.23解析实体测试数据

如图3.24当实体的真实（lst）和预测（ent\_pos）位置转化为二维向量后，可以得到实体位置是5，分类为1，而分类为1的真实值为1、2、9、10、11、12，即第一次判断错误。

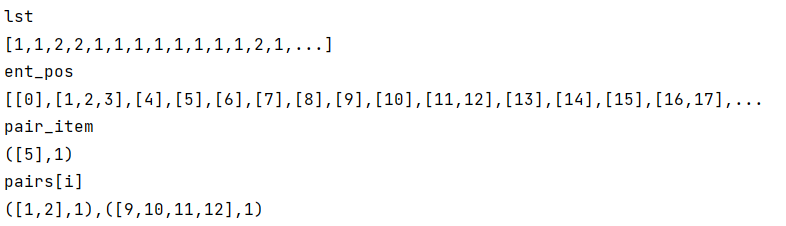


图3.24实体预测

核心代码如下图3.25：

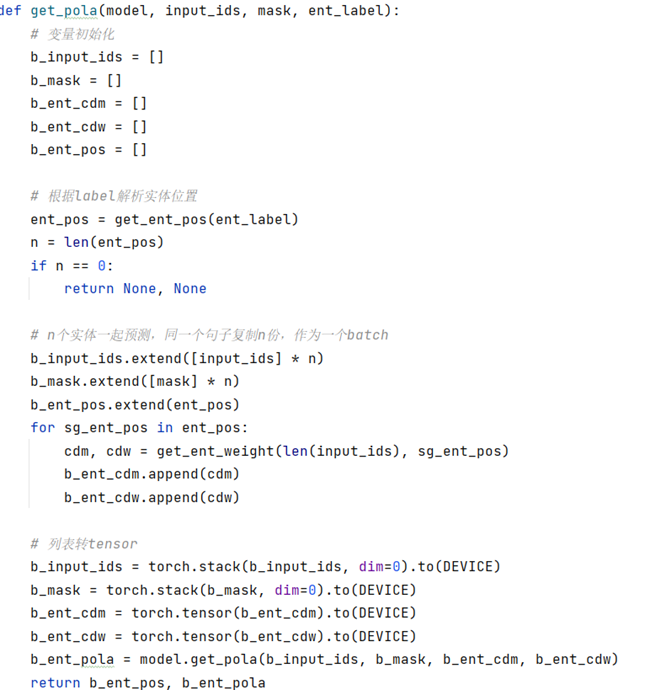


图3.25实体预测核心代码

## 3.9本章小结

本章我们完成了面向中文及多语言的LCF-ATEPC模型的具体设计与实现。首先对数据进行预处理，规范了数据的格式，剔除了异常值，然后对Bert词向量的获取，模型构建，模型训练，实体和情感预测的设计和思想做了详细介绍。

# 4 实验及结果分析

前面的章节介绍了本次实验的理论细节及实验原理，并且实现了实体标签和情感标签的预测，本章将介绍体现模型效率的各项评估指标。

## 4.1评估指标

### 4.1.1评估指标介绍

准确率（precision，正确率），定义为所有预测正确的样本除以预测总数，通常用于评估分类模型的性能，准确率越高说明模型分类的准确性越好。记为（correct\_cnt/pred\_cnt）。

召回率（Recall）是指在所有真实正样本中，被模型正确预测为正样本的样本数占总样本数的比例。用来评估模型在识别正样本时的表现，它越高表示模型在识别正样本方面的表现越好。记为（correct\_cnt/gold\_cnt）。

F1值是一个评估分类器或模型准确性的指标，它是准确率和召回率的平均值。F1值越高，表示分类器的准确性越高。F1值的取值范围是0到1，1表示最好的性能。记为（）。

### 4.1.2实验结果

为了避免效率过低，采用每一百次评估一次的策略。由于此项目为联合任务，所以实体和情感判断都正确才算对。首先要累加真实值数量，根据预测的实体label，解析出实体位置，并预测情感分类，即把实体，input\_id，label标签传入模型，解析实体和情感，并和真实值对比，在累加数值后，计算准确率，召回率，F1值。如图4.1可以看到当模型第五十次训练结束时，各项指标均已达到1，损失值也下降到较低位置，模型基本训练完毕。



图4.1模型训练情况

核心代码如下图4.2：

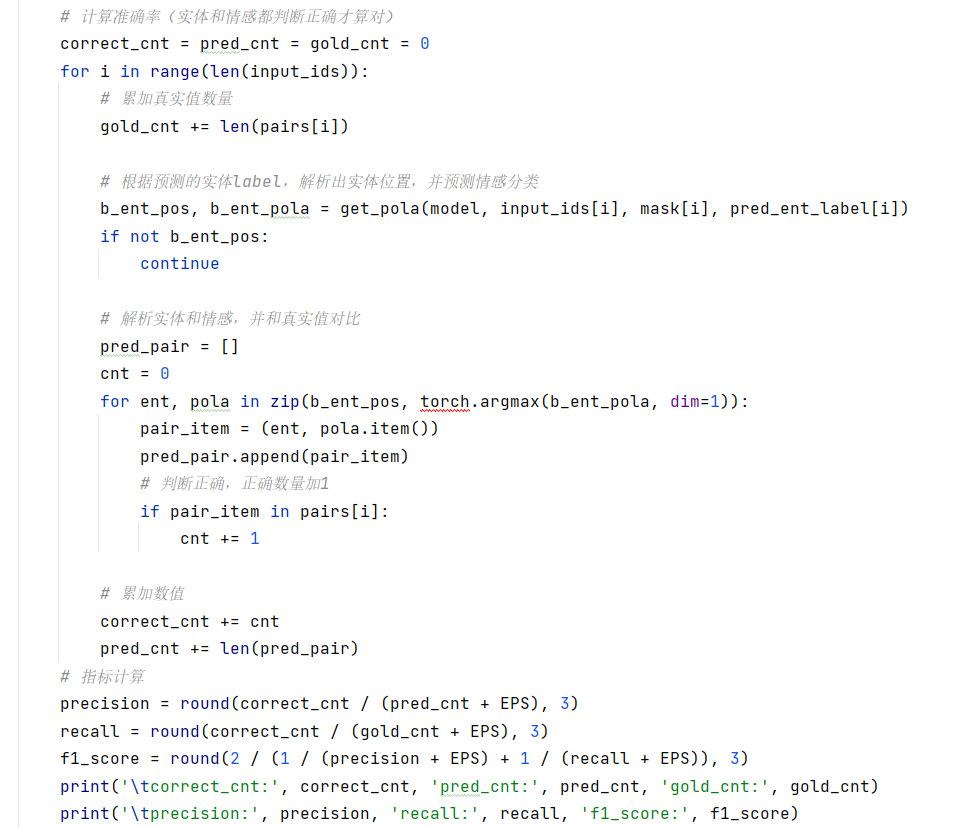


图4.2评估指标

## 4.2模型测试

完成模型训练并保存后，需要将数据集输入训练好的模型，查看模型最终效率。如图4.3经过25次循环后，可以看到模型最终准确率为75.1%。然后将“这个手机外观时尚，美中不足的是拍照像素低”这句话输入模型，经过tokenizer分词、encode转化成ID、实体和情感预测，得到前半句实体是“外观”，情感是“积极”，后半句实体是“拍照像素”，情感是“消极”判断正确。

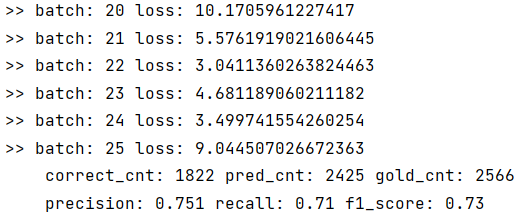




图4.3模型训练最终效果

## 4.3本章小结

本章主要介绍了体现模型效率的指标（准确率，召回率，F1值）的概念和计算方法，并对训练好的模型进行了测试。

# 5 总结与展望

本章将主要对基于LCF-ATEPC模型短文本情感分析在研究、设计及实现过程中应用或涉及的理论、技术和设计进行总结，并结合实际情况，对实验未来的发展、研究方向进行进一步的讨论。

## 5.1总结

在基于LCF-ATEPC模型短文本情感分析的研究、设计与实现过程中，应用或涉及了很多先进或是当今主流的技术和理论。

该模型具有动态特征加权和多头自注意力机制，可以更好地处理文本序列中的局部和全局上下文信息。还采用了基于CRF层的序列标注方法，有效地提高了模型的准确率和召回率。同时，在BERT层使用了预训练模型，提高了模型对评论文本的情感分析能力。通过在大量数据集上的实验，证明了 LCF-ATEPC 模型在情感分类任务上的有效性和性能卓越性。

## 5.2展望

使用深度学习和注意力机制解决自然语言处理问题已成为趋势，本文通过设计合理的模型结构，成功实现了基于LCF-ATEPC模型的短文本情感分类任务。尽管该模型已经有一定的效果，但仍有改进空间，需要在今后的研究工作中进一步优化和完善，具体内容如下：

（1）本文的研究重点是基于文本的情感分类，然而当前主流社交媒体平台中不仅包含用户的文本评论，还有大量的图片、视频等多媒体形式的内容。因此，未来的研究方向将探索如何将文本、图片、视频等多种媒体形式进行有效整合，从多个角度深入挖掘丰富的信息，综合分析用户的情感倾向。

（2）本文中采用了词向量表示作为文本表示方法。然而，现今越来越多的研究内容专注于基于文本的字向量表示和句向量表示。因此，在未来的研究中，可以通过比较这三种表示方法来找到最佳的表示方法，以提高模型的性能。

# 参考文献

1. H Yang，B Zeng，JH Yang，Y Song，R Xu, A Multi-task Learning Model for Chinese-oriented Aspect Polarity Classification and Aspect Term Extraction
2. Hu Xu, Bing Liu, Lei Shu, Philip S. Yu, BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis
3. Denver, Colorado,Proceedings of the 9th International Workshop on Semantic Evaluation(SemEval 2015)
4. 杨立公, 朱俭, 汤世平. 文本情感分析综述[J]. 计算机应用, 2013, 33(6):1574-1607
5. Tianyu Zhao,Junping Du,Zhe Xu,Ang Li,Zeli Guan, Aspect-Based Sentiment Analysis using Local Context Focus Mechanism with DeBERTa
6. Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis（ACL2020）
7. J Cordonnier, A Loukas, M Jaggi [EPFL], Multi-Head Attention: Collaborate Instead of Concatenate (2020)
8. Poria, S., Cambria, E., Ku, L.W., Gui, C., Gelbukh, A., 2014. A rule-based approach to aspect extraction from product reviews
9. Tang, D., Qin, B., Feng, X., Liu, T., 2016. Effective LSTMs for target-dependent sentiment classification
10. Devlin, Jacob，Chang, Ming-Wei，Lee, Kenton，Toutanova, Kristina, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

# 致 谢

在完成本篇论文的过程中，我获得了许多人的帮助和支持。

首先，我要感谢我的父母，感谢他们的坚定支持，为我提供舒适的生活，使我完成大学四年的学业。

然后，我要感谢罗鸿雁老师，因为在我的毕业设计过程中，她给予了耐心细致的指导，并且关注着实验过程中的每一个细节。在整个毕业设计的过程中，给予了我许多珍贵的意见和建议，让我在不断完善论文的过程中获得了很大的收获。还要感谢在新驰展娜科技有限公司实习期间的指导老师，范书维老师和朱晓琳老师，让我在实习中学习到宝贵的经验。

其次，我要感谢我的三个舍友：郑荣凯、郭洛林、张禾，他们是阿里云大数据应用学院数据科学与大数据计算专业2019级的学生。我们一起度过了一段美好的时光，互相扶持，相互帮助，形成了一种融洽友好的宿舍氛围。

同时，我也要感谢我所在的学校珠海科技学院，它提供了优秀的学术环境和学习资源，为我提供了一个良好的学习和研究平台。

最后，我还要感谢所有提供参考文献和资料的相关作者和机构，他们的贡献使得我能够更好地理解和探索研究课题。

至此，我再次向所有支持和帮助过我的人们致以深深的谢意。