《面向复杂分类体系的多标签情绪分类算法设计与实现》

本课题需要一定的自然语言处理(natural language processing)知识和机器学习(machine learning)基础。

具体来说，需要对自然语言处理领域中的情绪分析(emotion analysis)任务有一定的了解，并能够使用机器学习方法或其他方法解决相关问题。

编程语言方面以python为最佳(因为相应的第三方toolkit比较丰富和全面)。使用其他高级语言原则上亦无问题，不过会增加不必要的工作量。

任务背景：

在自然语言处理领域中，情绪分析任务旨在通过算法或模型来判断文本中所蕴含的人类情绪。一般来说，情绪分析的主流方法是先将情绪定义为若干个情绪类别，再将文本通过各种方法分到各自的类别中去。同时，由于人类情绪表达的复杂性与多变性，给定一个文本，其中可能包含一个或多个的不同情绪类别。而针对这一情况的情绪分析任务就称为**多标签情绪分类任务。**

目前不同情绪分类语料库往往采用不同的情绪分类体系，有粗粒度的体系(如仅将情绪定义为“正面”和“负面”两类)，亦有细粒度的情绪(将情绪定义为数十个不同的类型)。从应用的角度来说，人们总是希望分类体系的粒度越细越好(因为可以更加准确地描述文本中的情绪)。但是，由于目前的情绪分类方法多为单层次的方法(即认为各个情绪类别之间是独立的)，而这种方法对在情绪类别数量较小的情况下效果很好，但是在情绪类别较多的情况下往往出现分类准确率不高的情况。因此，本课题从不同情绪之间的层次关系出发，设计一种针对复杂情绪分类体系的多标签情绪分类的算法，以期能够获得比单层次分类方法更高的分类准确率。

数据集与课题任务：

本课题采用GoEmotions数据集[1]，系一个采用28类情绪类别体系的数据集。具体信息可以查看数据集压缩包里我写的说明文档。

该数据集的分类体系可以分为三个层次，粒度由粗到细。因此，可以从这三个层次的体系出发，先进行粗粒度的分类，再分别进行下一层的细粒度分类。例如，给定一个文本t，先对其进行最高层次的[A, B, C]三分类。如分为B类，则对再B类下的[c, d, e, f, g]五个小类进行判别，得到e、g两类。以此类推。

多标签分类策略：传统的多标签分类方法是将一个多标签分类任务建模为若干个单标签分类任务。以上例B类下[c, d, e, f, g]的五分类为例，一般方法是通过训练5个二元分类器，对每个类别进行true/false的二元分类。该方法称为二元相关法(Binary Relevance)。本课题在多标签分类的策略上基本以Binary Relevance即可。

[1] Dorottya Demszky,.Dana Movshovitz-Attias,et al. GoEmotions: A Dataset of Emotions[C]. Fine-Grained Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020.4040–4054.

实验方法与评估指标：

数据集已经划分为训练集、测试集和开发集。训练集中的所有信息(文本和类别信息)已知，测试集中假设只知道文本信息。利用训练集中的信息设计、训练相关的方法/模型，在测试集上验证方法/模型的有效性。开发集用于调节模型参数。有效性的高低根据评估指标来衡量。

多标签分类任务的评估指标一般采用准确率(Accurary)与F1值来评估，其计算公式如下：

手机屏幕的截图

描述已自动生成

图形用户界面

描述已自动生成

其中n为被测试的样本总数，是样本的真实类别标签集合(如样本t的真实标签为[d, f, g]，则该集合为{d, f, g})。是通过分类模型预测得到的标签集合(如样本t的真实标签为[e, g]，则该集合为{e, g})。|·|用于返回一个集合的大小。

本课题通过上述指标，通过对比所设计的多层次分类方法的性能与只使用单层次分类方法的性能，来验证方法的有效性。（注意这个指标和普通的acc以及F1有点区别）