文本分类已经成为从海量的数据中筛选出有价值的特定数据的一项重要技术。有效高效的文本分类方法,可以正确归纳文本主题，充分利用数据的价值，满足商业需求。文本分类技术已经被广泛地应用在文本审核、广告过滤、情感分析等领域。

最初的文本分类是通过人工审核的方法，费时费力。然后进化到通过抓取文本关键词，通过堆和树等数据结构进行简单的分类操作。这个方法依赖专家系统，需要专门构建特征规则，而且运行效率低且准确率低。随着人工智能的兴起，机器学习方法被用来进行文本分类任务。专家们利用专业知识，创建合适的输入特征，通过算法拟合出最优的模型作为输出，最后寻找文本中的某些特征，根据这些特征通过模型分为不同的类别。提升了准确率，但特征工程和计算能力限制了机器学习模型的表现。为了解决以上的问题，计算机科学家创造了深度学习方法，使神经网络自动通过每一层产生适当的特征，在实际应用中表现出色。如ELMo，OpenAI GPT，MT-DNN，XLNet，RoBERTa，SpanBERT等。本文研究的是基于BERT模型实现文本分类任务，并尝试了不同的方法优化长文本分类的结果。

BERT就是以带有Multi-head Self-Attention机制的Transformer结构为基本元素，通过预训练和微调的方法训练出的模型。BERT模型通过对上下文联合作用，来对深度双向表示的未标记文本进行预先训练。输入文本的嵌入由三种嵌入：Token Embeddings，Segment Embeddings和Position Embeddings求和而成。预训练分为两个任务，Masked Language Model和 Next Sentence Prediction。Masked Language Model在预训练时随机屏蔽了一部分的输入单词，然后预测那些被屏蔽的单词，为的是获得双向的预训练模型。Next Sentence Prediction 任务预测下一个句子是否是真正的下一个句子，为的是培养模型了解句子关系的能力。经过预训练的BERT模型可以通过一个额外输出层进行微调，以创造出适用于其他任务的模型。

本课题实验采用的是BERT-Base的中文版本，THUCNEWS中文新闻长文本标注数据集和今日头条中文新闻短文本分类数据集。对于短文本，我们只需要分类ID和新闻字符串。对于长文本，由于BERT模型有最长样例长度，为512个字，还包含开头的[CLS]和结尾的[SEP]，因此实际样例长度最多为510个字。为了解决这一问题，我们提出了以下几种思路：

（1） 取文本的前510个字（前截断）

（2） 取文本的后510个字（后截断）

（3） 取文本的前255和后255个字拼接（前后截断）

（4） 采用文本提取方法提取出长度适合的文本（文本摘要）

（5） 把超过510个字的文本拆分成多个长度合适的文本（拆分法）

截断的方法很好理解，但是这些截断方法，都会丢失一部分序列信息。方法（4）就是文本摘要法，此方法关键点是如何筛选出有效句子和有效信息。自然语言处理（NLP）中有两种文本摘要生成方法：抽取式和生成式。抽取式摘要通常需要衡量基本句子成分的权重，并根据权重结果生成摘要。生成式摘要的方法是，使用深度学习方法用于解释和缩写原始文档，生成表示源文本中最重要信息的新短语和句子。本课题采用的是抽取式摘要中的TextRank算法，通过把文本分割成若干组成单元，构建节点连接图，用句子之间的相似度作为边的权重，通过循环迭代计算句子的TextRank值，最后抽取排名高的句子组合成文本摘要。方法（5）是文本拆分，将一个整段的文本拆分为多个部分，每一个部分的长度小于510字。

由于我们直接使用已经训练好的BERT预训练模型，我们只需要进行微调过程。微调过程分为四个阶段：训练阶段，评估阶段，测试阶段和导出阶段。训练阶段进行输入文本的处理，模型的建立和样例的学习，评估阶段和测试阶段进行预测，导出阶段把模型导出为需要的格式。

经过实验，短文本单标签分类准确率最高为89.78%，达到了目前BERT-Base模型的基准值，也达到了目前主流文本分类模型的平均水平，相较于表现最好的模型，准确率相差不超过0.05% ，完全可以在实际的舆情系统中进行应用，达到了课题要求的标准。长文本单标签分类使用文本前后截取的方法效果最好，准确率最高为96.02%，比基准值高了0.67个百分点。这样的提升是因为前后截断非常适用于新闻类总分式的文本，前截断和后截断的方法处理总分式的文本表现都不够好。虽然拆分的方式没有截断法准确率高，但是文本拆分所保留了最完整的文本信息，处理文本的损失最少。而文本摘要的方法，由于在摘要的长度（要保留的句子数目）上不容易确定一个固定的数值，所以得到的摘要后的文本长度受单句长度影响较大，造成的后果就是文本长度不稳定，造成预测准确率偏低。

本课题还进行了基于英文多标签数据集的多标签文本分类实验。在多标签分类问题中，每个实例可以被分配到一组目标标签的多个类别，多标签分类与单标签分类主要区别有两个：第一，对象中的label属性是作为一个数组输入的；第二，我们使用sigmoid函数代替softmax函数，因为sigmoid函数会分别处理各个原始输出值，因此其结果相互独立，而Softmax函数的输出值相互关联。与单标签实验按照同样的方法进行训练。经过实验，多标签文本分类准确率为87.86%，与单标签文本分类差距较大。原因可能有两个：第一是多标签的文本属性导致多标签多类分类比单标签分类难度更大，第二是多标签数据集中的文本长度相对较短。

最后，本课题还在文本分类的基础上，加入了敏感词检测模块，使用DFA算法，以敏感词库中的词汇建立一个确定性的树形有限状态机，以循环的方式查询。如匹配，则用长度与敏感词长度相同的连续占位符代替原有的敏感词，并把处理好的文本返回。

在这些技术的基础上，本课题在已有的舆情系统上把上述功能进行集成。舆情平台的主要功能包括：实时获取信息平台的舆情信息、快速检索舆情信息文本、快速发现舆情热点并进行分析与预警，生成舆情报表使用户更直观全面了解舆情信息等。舆情系统的工作流程可以划分为五个阶段：获取数据、处理数据、存储数据、逻辑业务和用户交互。，通过各种网络爬虫进行原始数据采集，通过Kafka对数据进行初步处理，存储到不同的数据库，业务逻辑层负责将数据存储层的数据处理成为用户需要的特定格式，可视化的展示给用户。

本课题在业务逻辑层把文本分类和敏感词检测功能进行封装，实现了舆情系统的一条完整的业务流程，证明了上文中训练的文本分类模型和敏感词检测模块的有效性。