

# 上海交通大学

SHANGHAI JIAO TONG UNIVERSITY

## 本科生毕业设计（论文）开题报告



论文题目: 舆情文本多分类与敏感性检测技术研究与实现

学生姓名: 周笑

学生学号: 516030910370

专    业: 软件工程

指导教师: 唐新怀

学院 (系): 电子信息与电气工程学院

教务处制表

## 填表说明

1. 根据《上海交通大学关于本科生毕业设计(论文)工作的若干规定》要求,每位学生必须认真撰写《毕业设计(论文)开题报告》。
2. 每位学生应在指导教师的指导下认真、实事求是地填写各项内容。文字表达要明确、严谨,语句通顺,条理清晰。外来语要同时用原文和中文表达,第一次出现的缩写词,须注出全称。
3. 开题前,须进行文献查阅,要求与论文研究有关的主要参考文献阅读数量不少于 10 篇,其中外文资料应占一定比例。参考文献的书写请参照《上海交通大学本科生毕业设计(论文)撰写规范》。
4. 毕业设计(论文)开题报告总字数应满足本院(系)要求。
5. 请用宋体小四号字体填写,并用 A4 纸打印,于左侧装订成册。
6. 该表填写完毕后,须请指导教师审核,并签署意见。
7. 《上海交通大学本科生毕业设计(论文)开题报告》将作为答辩资格审查的主要材料之一。
8. 本表格不够可自行扩页。

## 毕业设计(论文)开题报告

论文题目	舆情文本多分类与敏感性检测技术研究实现				
课题来源	预研	课题性质	设计	项目编号	
<b>课题研究目的和意义（含国内外研究现状综述）：</b> <p>文本分类是指计算机将载有信息的一篇文本映射到预先给定的某一类别或某几类别主题的过程。随着大数据时代的到来，文本分类正越来越受到大家的关注。有效的文本分类方法，可以建立智能推荐系统，使其可以根据用户的个人兴趣来定位并推荐相关的新闻资料。这样的优势使得文本多分类任务愈发受到企业的青睐。</p> <p>此课题就基于 BERT<sup>[9]</sup>（Bidirectional Encoder Representation from Transformers）方法来实现文本多分类的功能。BERT 诞生于 2018 年 10 月，作为当时最新的 state of the art 模型，通过预训练和精调横扫了 11 项 NLP 任务，相对 RNN 更加高效、能捕捉更长距离的依赖。SQuAD 2.0 leaderboard<sup>[10]</sup>中前 17 个系统，CoQA leaderboard<sup>[11]</sup>排前五的系统全部是基于 BERT 模型设计的。这足以说明 BERT 模型在文本分类中的能力。</p> <p>国内外对于文本多分类也进行了长时间的研究，获得了许多成果。比如 XLNet<sup>[1]</sup>模型，在许多方面达到了超越 BERT 的表现。XLNet 还提出了新的训练机制，如 Permutation Language Model，Two-Stream Self-Attention 和 Recurrence Mechanism<sup>[2]</sup>。此外，还有其他模型提出基于 BERT 的改进，让 BERT 发挥更大的潜能。如 Facebook AI 联合 UW 发布的 BERT 预训练模型 RoBERTa<sup>[5]</sup>，通过修改模型的预训练任务和目标使模型达到更好的效果的 SpanBERT<sup>[6]</sup>，以及将 Multi-Task 与 BERT 结合起来，使得模型能在更多的数据上进行训练的同时还能获得更好的迁移能力的 MT-DNN<sup>[7]</sup>。</p> <p>此外，还有学者尝试在不同情境下提高 BERT 模型的表现。如 Argument Reasoning Comprehension<sup>[3]</sup>任务和 Natural Language Inference<sup>[4]</sup>任务，都取得了显著的提高<sup>[8]</sup>。</p> <p>虽然 BERT 在许多多面仍有提高的空间，但 BERT 作为这些新模型的基础，仍具有可观的研究价值和提升潜力。此课题旨在通过 BERT 方法，实现在真实平台上可用，有效，高效的文本多分类模型，并在敏感词检测任务中进行实践。</p>					

**课题研究内容：**

主要任务是结合 BERT (Bidirectional Encoder Representation from Transformers) 预训练模型对舆情内容进行多分类。研究如何结合 BERT 技术检测中文舆情文本敏感性并提升中文舆情内容多分类准确率。将相关研究成果应用于舆情事件监控领域。

**研究方法和研究思路（技术路线）：**

根据课题需求，我把此课题的研究分为三部分：模型训练部分，敏感词检测部分和集成部分。

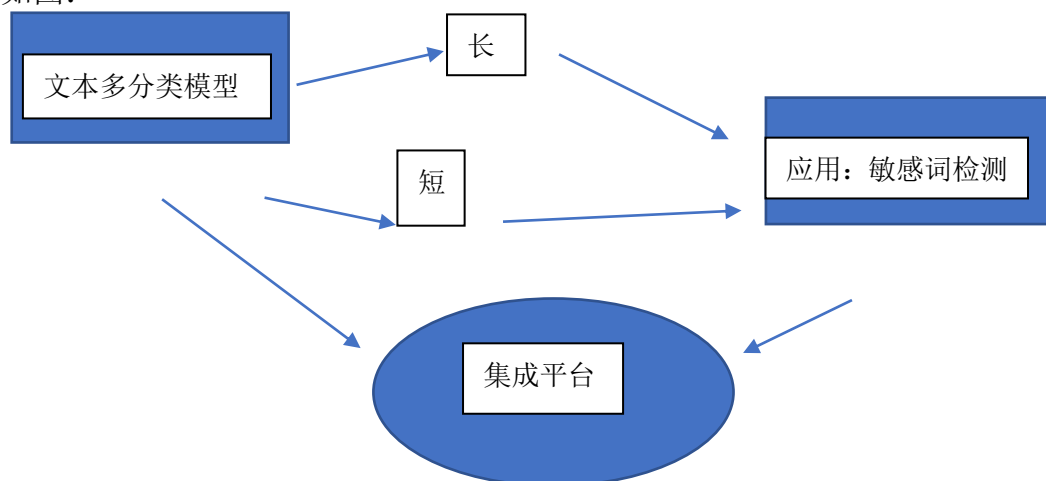
首先我会根据文本长度，将文本分类为长文本和短文本，并根据长短文本不同的特性，对训练集和训练模型进行不同的参数设置，并在测试模型的时候根据表现进行调整。实现针对长或短文本的针对性训练模型。目的是实现更好的准确度，以达到课题的要求。

第二部分是敏感词检测。相对于第一部分来说，敏感词检测相当于一个文本多分类的应用。但是数据集与普通长短文本不同。因此，我会根据敏感词检测的要求，调整训练集和训练方式，以求达到最大化的准确度和效率。

第三部分是集成。因为我们的模型最终都要在集成平台上使用，因此能否在集成平台上实现训练集中的效果也是一项艰巨的挑战。因此我会多加测试，并把集成环境中的结果与训练结果相比对，从中发现问题，并作出针对性的调整，使我的模型能够很好地兼容平台，并做到准确度达标。

综上所述，文本多分类模型是敏感词检测的基础，敏感词检测是一项文本多分类的应用，而这两者都要受到集成平台的制约和影响。我将把重点放在这三部分的实现上，以期高品质地完成这个课题。

如图：



**预期研究结果：(可选填)**

- (1) 研究基于评价指标准确率 (accuracy) 的中文舆情内容长文本分类, 基于 THUCNEWS 中文新闻长文本标注数据集 ([http://106.13.187.75:8003/download/?dataset\\_name=thucnews](http://106.13.187.75:8003/download/?dataset_name=thucnews)) 的准确率不低于 95.35
- (2) 研究基于评价指标准确率 (accuracy) 的中文舆情内容短文本分类, 基于今日头条中文新闻短文本分类数据集 ([http://106.13.187.75:8003/download/?dataset\\_name=tnews](http://106.13.187.75:8003/download/?dataset_name=tnews)) 的准确率不低于 89.78
- (3) 研究基于中文舆情文本关键词、语义的敏感内容检测技术。
- (4) 在现有舆情事件分析系统中集成上述方法。

**计划进度安排：**

- （1） 在第一阶段检测时，实现针对中文舆情内容实现结合 BERT 技术的长短文本的分类模型训练。
- （2） 在中期检测时，实现针对中文舆情内容实现敏感性检测。
- （3） 在五月完成本科毕业论文，论文不少于 2.5 万字。

**参考文献:**

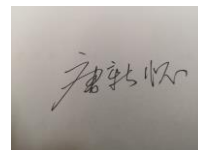
- [1] XLNet: Generalized Autoregressive Pretraining for Language Understanding. Yang et al. CoRR abs/1906.08237.
- [2] A Fair Comparison Study of XLNet and BERT. XLNet Team. <https://medium.com/@xlnet.team/a-fair-comparison-study-of-xlnet-and-bert-with-large-models-5a4257f59dc0>
- [3] Probing Neural Network Comprehension of Natural Language Arguments. Niven et al. ACL2019.
- [4] Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. McCoy et al. CoRR abs/1902.01007.
- [5] RoBERTa: A Robustly Optimized BERT Pretraining Approach. Liu et al. CoRR abs/190.11692.
- [6] SpanBERT: Improving Pre-training by Representing and Predicting Spans. Joshi et al. CoRRabs/1907.10529.
- [7] Multi-Task Deep Neural Networks for Natural Language Understanding. Liu et al. CoRR abs/1901.11504.
- [8] Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding. Liu et al. CoRR abs/1904.09482.
- [9] Bert: Pre-training of deepbidirectional transformers for language understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. arXiv preprint arXiv:1810.04805
- [10] Know what you don't know: Unanswerable questions for squad. Pranav Rajpurkar, Robin Jia, and Percy Liang. arXiv preprint arXiv:1806.03822
- [11] Coqa: A conversational question answering challenge. Siva Reddy, Danqi Chen, and Christopher D Manning. arXiv preprint arXiv:1808.07042



指导教师意见（课题难度是否适中、工作量是否饱满、进度安排是否合理、工作条件是否具备、是否同意开题等）：

课题难度较大，工作量满足本科毕业设计要求，进度安排合理，实验室满足相关研究条件。

指导教师签名：\_\_\_\_\_



2020 年 3 月 23 日

学院（系）意见：  
审核通过

审 查 结 果： ☒ 同 意    ☐ 不 同 意

学院（系）负责人签名：\_\_\_\_\_

2020 年 3 月 23 日



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY

毕业设计（论文）开题报告

---