各位老师好，我是周笑，我的毕业设计的课题题目是舆情文本多分类与敏感性检测技术研究与实现。我的PPT陈述环节分为五个部分，首先介绍课题，然后介绍模型和算法，然后介绍实验的设计和结果分析，最后在舆情系统中实现文本分类和敏感词检测功能。

首先我们都知道，移动互联网已经在人们生活中普及，微博，微信，qq，许许多多的平台每天都会产生大量的舆论信息，这些舆情信息代表了人们普遍看法和认识，是民众思想状况的一种反映方式。将人们的社会政治态度、看法和意见进行收集、整理，是新形势下的网络安全工作的需要，因此，可以对网络上各个平台的各类信息进行采集、分类、整合、筛选的舆情系统也就应运而生。而且，随着大数据时代到来，数据的体量越来越大，如何从海量的数据中筛选出有价值的可以满足需求的特定数据，并使其发挥出本身的价值，文本分类就是解决这一问题的一项重要技术。

本课题的研究目的，是研究更准确的文本分类模型&更高效的敏感词检测方法，并把分类模型和检测算法集成到舆情系统中去，实现文本分类和敏感词检测两个功能模块。

文本分类方法从最初的人工分类，到基于规则的匹配，例如篮球映射到体育，电影映射到娱乐等。这些方法需要专门构建特征规则，费时费力，效率低，准确率也低。随着人工智能的发展，出现了许多机器学习方法，比如朴素贝叶斯分类，决策树等，后来又出现了深度学习方法，cnn，rnn还有lstm长短期记忆模型。本课题采用的BERT模型就是最新的主流的深度学习方法。

敏感词检测方法也是从人工检测，到字符串匹配，再到机器学习方法。但是作为一个舆情系统，准确率的要求是100%，当前的机器学习方法不能满足准确率的要求，所以目前主流的方法是字符匹配。主要是DFA算法和AC自动机算法。

接下来我介绍本课题所依靠的舆情系统，结构分为五层，首先是数据获取，通过正文抽取和scrapy爬虫框架，从网页和微博等平台上获取数据，然后把数据送入kafka流处理平台，就是数据处理层，把上一层获取的数据进行处理，方便后面的业务流程进行调用。数据存储层使用HBase等数据库，分别存储不同类型的数据。业务逻辑层就是从数据库中获得数据进行统计，检索等流程，送入前端用户交互层展示。我的课题就是向数据处理层中添加文本分类模块和敏感词检测模块，实现业务流程。

本课题文本分类任务使用的模型是BERT模型，是2018年9月google提出的。BERT最大的特点在于双向，同时联系上下文的信息。这是BERT相比其他模型最大的进步。而且BERT使用预训练+微调的方法，就是迁移学习的方法，使得模型具有很好的泛化能力。

为什么说BERT是双向的？首先，BERT的基本组成元素是双向的Transformer。在传统的神经网络中，输入数据经过一层层编码器，得到一个上下文向量，把上下文向量作为解码器的输入，下一层解码器的输入只有上一层解码器的输出。而transformer采用attention机制，把每一层编码器的输出都保留并输入到解码器中，解码器通过计算每一层编码器输出的attetion值，就是每一层编码器的输出重要程度，选择最重要的编码器输出作为输入进行解码。这样就可以把数据中的重要信息进行充分学习。BERT模型采用的是双向的multi-head self-attention机制。Self-attention就是学习一个单词的上文对于该单词的重要程度。Multi-head就是把输入进行多次投影再进行计算，这样可以使模型能够更加关注输入数据中的重要信息。而双向，代表BERT模型的transformer可以同时学习到上文和下文的信息。

BERT的双向还体现在预训练任务中。完形填空任务就是为了训练双向的语言模型。。。