大家好，我的项目是“基于半监督学习的弹幕情感分析与过滤”。

首先说明一下，我所说的弹幕是指看视频时大量飘过的这种弹幕，不是微信墙上的那种弹幕。

项目重要性，弹幕在青年群体的影响力非常大，B站作为中国最大的弹幕视频网站，截止2017年第四季度，活跃用户为7180万人，日均使用时长超过1小时，可见弹幕视频网站已经成为青年人的主流选择之一。而弹幕视频网站如此火爆，和弹幕本身的价值不无关系。现在中国社会的独生子女和城市独居青年较为普遍，他们更愿意在看视频时通过弹幕与人交流，获得陪伴感，缓解孤独。

但是弹幕现在也存在着一些问题，有人用弹幕刷屏影响观看，这个通过相近弹幕合并就很容易解决，而有人用弹幕吵架起来了，或是宣泄负能量，就需要通过情感分析来进行过滤。

弹幕的特点决定了方法：首先它是超短文本，常常不符合语法规范，所以现有很多的分析微博文本的方法不好直接搬用到弹幕上，需要上下文且要求预料准确的word2vec也不好用。

第二点是部分的词在弹幕中使用时含义发生变化，不是平时的用法，弹幕中还有许多新创的网络词汇。那么我们可以将这些词选出来作为带学习目标，而其他词认为是正常中文的用法。利用上现有的成熟的中文常用词情感词汇库，就可以去给这些其它词打标签。那么为什么我选择了半监督学习而不是监督学习呢？

来看看实际的数据情况，我爬取4万多条弹幕，通过TF-IDF选出了500个弹幕词。TF-IDF就是一个衡量一个词对一个语料库的重要程度的指标。可以看到弹幕词虽然只有500个，但却在这4万条弹幕里出现了约15000次，而其他在情感词汇库里有的词有3400个之多，却只出现了约5500次，可以看到大量带挖掘信息是存在于这500个弹幕词汇中的，这是少量有标签样本，而这是大量无标签样本，用半监督学习，正好。

说明一下打标签用的是大连理工信息检索研究室的中文情感词汇本体库，它是在国外比较有影响的Ekman的6大类情感分类体系的基础上，对“好”的情感又细分出两类得到的7维情感体系，这7维分别是快乐，赞扬，愤怒，悲伤，恐惧，厌恶，惊奇，这个库里有两万七千多个中文常用情感词的情感向量。

输入输出的处理：比如这有一条弹幕，用结巴分词后得到下面这些分开的词，其中阴谋和伤害这两个词在中文情感词汇库中，对应的情感分别为阴谋的厌恶值为5，伤害的悲伤值为3，那么这条弹幕的情感向量就是这个样子。而弹幕是有在视频内的发送时间的，于是就可以得到这个视频的7维情感随时间变化的信息。然后每一个词当作是一个sample，输入是这个词所在的弹幕的前后各10秒的7维情感，即是得到一个20\*7的矩阵输入，输出则是这个词的7维情感向量。

半监督学习的基本流程：首先用LSTM网络监督学习有标签样本得到一个网络模型，用这个网络去预测未标签样本得到其预测的标签，然后将未标签样本分成若干组，计算每组的置信度，将置信度最高的那组加入到有标签样本，重复这个循环直到达到训练阈值。

由于这是个回归问题，用不了半监督分类的那些置信度计算方法，但是本项目又和一般的半监督学习有区别，就是未标签样本的预测标签可以影响输入。所以我的办法是，将每组弹幕词的预测值当作真实值去更新有标签样本的输入，用原神经网络对更新后的输入进行预测，计算新的预测值和其原标签的均方差。均方差大则表明这组弹幕词的预测标签和原有的情感向量随时间变化的情感背景冲突大，认为其置信度低。

这是从最终的弹幕词典里选出的几个词，可以看到各维情感的强度值一般在0~7分布，且词的情感不一定是只在单一维度上有非零值，比如橘里橘气这个词，乐的情感强度为2，恶的情感强度为1.

最终训练好的网络在最初的有标签样本上的预测值和原标签是均方差0.7，在7维上有一个不到1的均方差说明网络效果还是很好的，确定系数0.9相当接近1也说明网络效果很好

然后看一下具体是如何使用的，首先找到一个你想看的视频，在网页上获得这个视频的编号，将其放到程序里运行，程序就会去爬取这个视频的弹幕进行处理，首先画出这个视频的7维情感随时间变化的曲线，同时会生成一个xml文件，里面是这个视频要被过滤掉的弹幕，将导入b站，看视频时就没有这些弹幕了。

得到视频情感曲线有利于观众在看视频前对视频情感有个大致预期，比如你现在心情不太好，看到这个视频厌恶情感值全程较高，而快乐情感值除了一个小高峰全程较低，你就知道现在不适合看这个视频。

这个是我做的弹幕过滤前后的效果对比图，上方是过滤前，下方是过滤后。

弹幕飘得太快不好分析，我截取刚才的一屏出来，可以看到，像你情敌死的时候呢，温柔根本毫无意义，只有痛苦和折磨才是人生的真谛这种负能量爆棚弹幕就被过滤掉了，而 泪目，还有这变化简直辜负了祭这些正常表达悲伤情感或讨论剧情的弹幕就被保留了下来，可见通过情感分析来过滤弹幕效果还是很好的。

我们希望通过情感分析过滤来让弹幕文化更好地发展，感谢观看！欢迎提问

项目完善上：预测情感值的模偏小——之前处理，现在

可继续研究的点——1.预测的值会影响输入的半监督学习

2.利用双时间轴信息区分表达欢乐情绪和引发欢乐情绪的弹幕