陆续训练了：

1. 数据不均衡/均衡下的，分别基于bert-base-chinese/electra-180g-small/finbert的，二分类/三分类模型，用于情感分类。

目前使用的是数据均衡下，基于finbert的三分类模型。

1. 数据不均衡/均衡下的，分别基于bert-base-chinese/electra-180g-small/finbert的，二分类模型，用于分类文本是否有效。广告、招聘、销售等无效文本会被过滤。

目前使用的是数据均衡下，基于finbert的二分类模型。

1. 基于finbert/gpt2-base/gpt2-small的中文命名实体识别模型。

目前使用的是基于finbert的NER模型。

1. 基于finbert的二分类模型，用于区分那些和主体相关但是无效的文本，例如“渤海银行门前有人撞车”等。

目前该模型未被使用，原因是训练语料不充分，模型性能有待提高。积累预料之后，可以替换2和3。

目前的分类流程是：

文本🡪3.NER模型🡪2.有效无效🡪1.情感分类

使用的代码在app2.0文件夹中

目前调研了1：n对比文本相似度的算法；正在调研中文分词算法。

接下来的工作：

1. 加速当前流程；
2. 识别并剔除转发的文本；
3. 新闻预警级别分析；s