邓娜代码分析

事件分析

feature_extraction.py

利用卡方检验来判断某个词是不是情感词——即该词与正负情感有没有关系。

https://www.jianshu.com/p/807b2c2bfd9b

ChiSquare类

- `__init__()`：利用文档标签和文档单词构造总"单词-频数”字典、正"单词-频数”字典、负"单词-频数”字典。
- `__calculate()`：进行卡方检验，计算得到的卡方值作为特征得分pos_score。计算公式为：

$$X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

- `best_words`：按特征得分对单词从大到小排序，取出num个特征词。

Corpus.py

构建语料库。

Corpus类

- `__init__()`：读入包含 *带正负标签的文档* 的文件构造正负文档数组。
- `get_corpus()`：从正负文档数组中选取数量相同的数据（从start到end），返回数据和数据标签(1或0)。
- `get_train_corpus()`：调用 `get_corpus()` 函数，返回0到num这么多的数据及标签。
- `get_corpus()`：调用 `get_corpus()` 函数，返回train_num到train_num + num这么多的数据及标签。
- `get_all_corpus()`：返回所有数据和数据标签(1或0)。

派生类

以下类均是继承自Corpus类，初始化时指定了语料库文件位置：

MoiveCorpus类

Moive2Corpus类

WaimaiCorpus类

Waimai2Corpus类

HotelCorpus类

准备语料库相关函数

准备语料库txt：从本地的数据源得到数据，并在f_corpus中生成对应的xx_corpus.txt文件，便于后续使用。

`get_movie_corpus()`

`get_movie2_corpus()`

`get_hotel_corpus()`

`get_waimai_corpus()`

测试函数

`test_corpus()`：测试以上派生类。

classifiers.py

分类器。

DictClassifier类

- `__init__()`：准备用户词典、情感词典、其他词典。
- `classify()`：调用`analyse_sentence()`函数进行分类。
- `analysis_file()`：对整个文件分析。循环调用`analyse_sentence()`函数对单个句子分析。
- `analyse_sentence()`：将评论分句，对每个子句进行情感分析；根据最后得分的正负返回1或0.
- `__analyse_clause()`：判断各种句式，然后逐个分析分词判断类型，然后综合连词的情感值和标点符号的情感值。返回子句分析的数据结构。

```
sub_clause =
{
  "score": 0,
  "positive": [],
  "negative": [],
  "conjunction": [],
  "punctuation": [],
  "pattern": []
}
```

判断句式函数

- `__is_clause_pattern1()`
- `__is_clause_pattern2()`
- `__is_clause_pattern3()`

分析单词及判断词类型

- `__analyse_word()`：调用以下函数判断词的类型。

查找对应词典，看词是不是在词典中来判断类型。

- __is_word_conjunction
- __is_word_punctuation
- __is_word_positive
- __is_word_negative

情感词分析

`__emotional_word_analysis()`：在情感词典内，构建一个以情感词为中心的字典数据结构：

```
orientation =
{
  "key": core_word,
  "adverb": [],
  "denial": [],
  "value": value
}
```

在三个前视窗内，分别判断是否有否定词、副词、情感词，更新数据结构，添加情感分析值，最后返回该数据结构。

分句相关函数

- `__divide_sentence_into_clauses()`：先调用 `__split_sentence()` 根据标点符号分割句子，然后识别句式来分句。
- `__split_sentence()`：根据标点符号分割句子。

其他函数

- `__get_phrase_dict()`：读入短语词典。
- `__get_dict()`：读入词典。
- `__write_runout_file()`：写入文件操作。
- `__output_analysis()`：输出分析的数据结构结果。

KNNClassifier类

所谓K最近邻，就是k个最近的邻居的意思，说的是每个样本都可以用它最接近的k个邻居来代表。

kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。

- `__doc2vector()`：将文档中的单词转换为向量，向量维数为单词总数，每一维是对应单词的出现频率。

略。

BayesClassifier类

- `__train()`：输入训练数据，获得每个单词的频数和概率信息。
- `classify()`：用正负情感词的概率之和作为情感得分（分别计算pos_score和neg_score），如果pos_score大于neg_score则输出正向，否则输出反向。

MaxEntClassifier类

略。

SVMClassifier类

- `words2vector()`：词转化为词向量，向量值根据特征词的频数得到。
- `__train()`：输入训练向量，调用SVC的fit函数进行训练。
- `classify()`：词转化为词向量，调用SVC的predict函数进行预测。
- `predict_proba()`：调用SVC的predict_proba函数进行预测。
  - predict是训练后返回预测结果，是标签值。
  - predict_proba返回的是一个 n 行 k 列的数组， 第 i 行 第 j 列上的数值是模型预测 第 i 个预测样本为某个标签的概率，并且每一行的概率和为1。

tools.py

Write2File类

- `append(), write()`：将结果写入文件（普通文本）中
- `write_contents()`：将准确率写入xls表格中
- `write_results()`：将结果写入xls表格中

`get_accuracy()`：计算分析后标签的准确率（利用F-measure方法对比实际标签和预测标签）

test.py

测试：分别测试各个方法，各个语料库。

过程：

1. 获取训练、测试语料集
2. 特征提取
3. 基于不同方法进行分类
4. 计算准确率结果

技术框架总结

爬取数据

- 新闻文本：人民日报、搜狐新闻、澎湃新闻、腾讯新闻、食品伙伴网
- 新闻音频：蜻蜓FM、喜马拉雅 `来源可以修改`
- 新闻视频：央视网

数据处理

文本处理

- 数据清洗：删除带有杂质、重复的数据
- 中文分词
    - 使用 pkuseg 分词工具进行分词（分词准确率高、可支持不同领域的分词）
    - Jieba分词（文晴曼）
    - 采用 NLPIR 分词系统（细粒度）：用户词典功能，将识别出的新词添加到用户词典
- 去除停用词
    - 综合四川大学机器智能实验室停用词表、哈工大停用词表等
    - 百度停用词表（文晴曼）
- 特征选择和提取 `针对数据稀疏、内聚性差的数据`
    - （文晴曼）：利用TextRank算法选择重要性较高的词语，构建**主题词典**，利用TextRank算法提取每个文本的前 top-K 个关键词，再从这些关键词集合中选择出现次数最多的前 top-N 个词语，组合这些词语形成主题词典
    - （文本话题识别）：TF-IDF算法选取值最高的K个关键词，作为文档关键词

音频处理

- Python 的 imageio 库对各种图像数据进行读取和写入操作
- 利用 ffmpeg 批量转化音频格式（pcm、wav、amr）
- 百度语音识别API将音频转换为文字

视频处理

- Python 的 imageio 库对各种图像数据进行读取和写入操作
- 利用 ffmpeg 进行解码视频文件
- 利用**百度文字识别**工具对视频中的文字进行识别 `可能不需要`

文本向量化

- 使用神经网络模型中 word2vec 模型训练词向量
- 利用训练好的 word2vec 模型将输入的词语转化为词向量表示
- 将词向量按照文本顺序拼装，生成词向量矩阵
- 利用人民日报中文语料库，使用 gensim 中 word2vec 学习词向量
- 去除同义词 （文本话题识别） ：准备完备的同义词表，利用 Wikipedia 语料库训练 word2vec 模型用于对多义同义词归并进行确定

文本过滤

利用支持向量机 SVM

> 对文本预处理后的文档进行分类，过滤掉与食品安全无关的文本

- 人工标记数据集
  - 与食品安全相关：主题爬虫爬取和食品安全相关的新闻
  - 与食品安全无关：爬取与娱乐、体育相关的新闻数据
  - 按 7:3 划分训练集和测试集
- 将训练集输入训练分类器模型，测试集输入用来评估整个模型效果
  - 评价指标：混淆矩阵
- 对新闻文档进行判别，过滤掉与食品安全无关的文档

主题提取

LDA主题提取

> 给定一篇文档，通过训练好的 LDA 模型预测出该食品安全相关文档所隐藏的主题词集合

- 确定最优主题数T
  - 利用困惑度（使困惑度最小）
  - 利用贝叶斯统计标准方法
- 将构建的最优模型存储在文件中
- 新文档需要确定主题时，加载已保存的主题模型，预测文档的主题分布

话题分类

文本聚类

LDA+VSM建模

1. 读取语料文件，根据 gensim 库生成语料模型；

2. 生成 LDA 模型、VSM 模型并保存模型文件；
3. 使用两个模型对语料进行转换，转换为多维向量，其中 VSM 模型通过TFIDF 算法将文档转换为多维 VSM 向量， LDA 模型根据文档-主题的概率分布，生成一个各维相加为 1 的多维 LDA 向量；
4. 根据两种向量分别计算相似度，使用加权值对两个相似度进行整合计算，其中 VSM 向量使用余弦相似度计算向量的相似度，LDA 向量使用 JS 距离的方法计算向量的相似度；

LDA 相似度计算

> JS 距离

$$\mathrm{D}_{js}(p,q)=\frac{1}{2}\left[D_{KL}\left(p,\frac{p+q}{2}\right)+D_{KL}\left(q,\frac{p+q}{2}\right)\right]$$

VSM相似度计算

> 余弦相似度

- TF-IDF：特征项加权
- 余弦定理衡量相似度

文本聚类

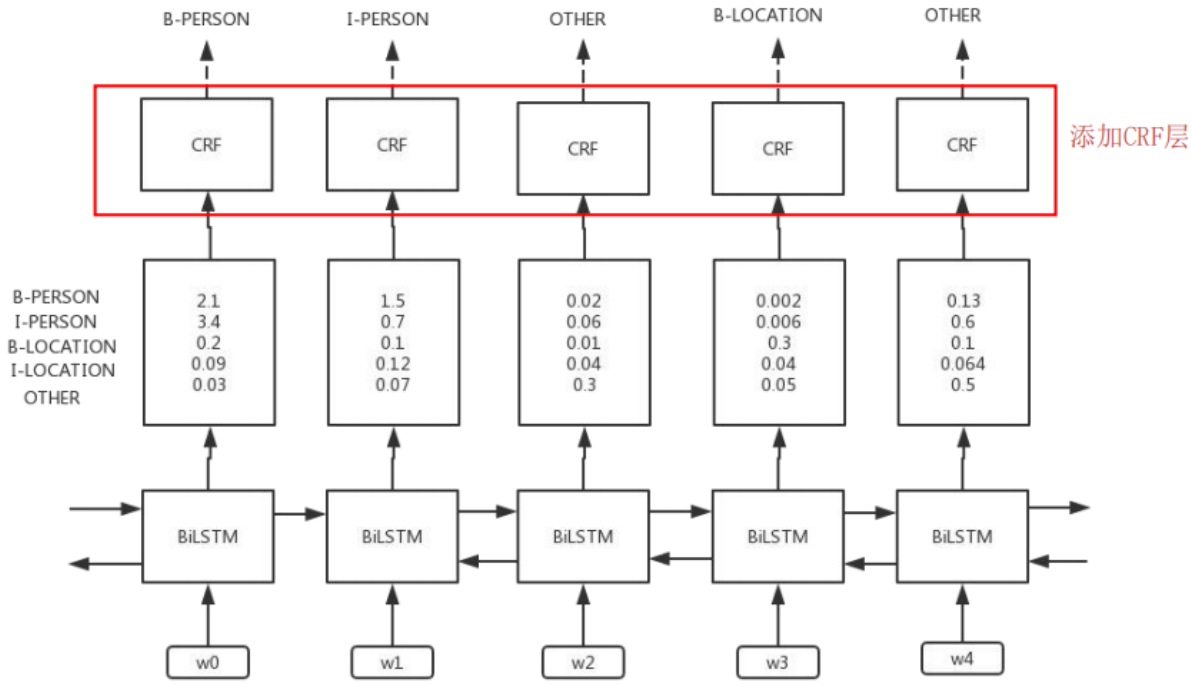
在相似度的基础上进行自底向上的层次聚类，返回每一个文档对应的类别数。

事件提取

特征项提取

> 提取新闻文档中食品安全专有名词、标记食品安全名词、人名、地名、机构名

- 采用BiSLTM-CRF模型对句子进行标注
    - PER（人名）、LOC（地名）、FOOD（食品安全专有名词）、ORG（机构名）



- 将训练好的模型存储

文本相似度计算

> 计算新文档与事件的相似度，用来判断该文档是否属于新事件

$$\mathrm{Sim}(N_1, N_2) = \theta_1 \,\mathrm{Sim}(x_1, x_2) + \theta_2 \,\mathrm{Sim}(y_1, y_2)$$

$$\mathrm{Sim}(X, Y) = 0.5 + 0.5 * \cos(X, Y)$$

$$\cos(\mathrm{X}, \mathrm{Y}) = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$

N：文本数据；x：食品专有名词向量；y：文本数据向量

- 若相似度小于阈值，则新增一个事件
- 若相似度大于阈值，则归为该事件之中

情感分析

文本卷积神经网络

- 结合食品安全现有语料库以及爬取文本建立的相关语料库
- 基于词向量和 TextCNN（文本卷积神经网络）进行情感倾向分析

词向量构建模块

- 数据处理：数据清洗、中文分词、停用词
- 采用 word2vec 模型中的 skip-gram 模型进行训练
    - 输入N维的one-hot向量

    - 隐含层的权重矩阵：词向量

分类器模块

> 文本卷积神经网络

- 输入层：输入词向量（文本序列中各词汇的词向量表示）
- 卷积层
    - 输入文本中各词对应的词向量矩阵
    - 窗口值：相邻该词前后N词的信息
- 池化层：最大值池化
    - 结合卷积层进行特征提取
    - 提取的特征值组成特征图
- 全连接层：类似于分类器，经过softmax激活函数，得到分类结果xili

图像文本联合情感分类

图像情感分类

> MFES：多级特征提取器

- 将输入图像和每一个卷积层的输出特征图分别输入一个 MFES，MFES 会分别生成一个一维向量，将这些输出的向量(即 e2，e3，e4，e5，e6)与原卷积神经网络的输出向量 e1 拼接，作为全连接神经网络的输入特征

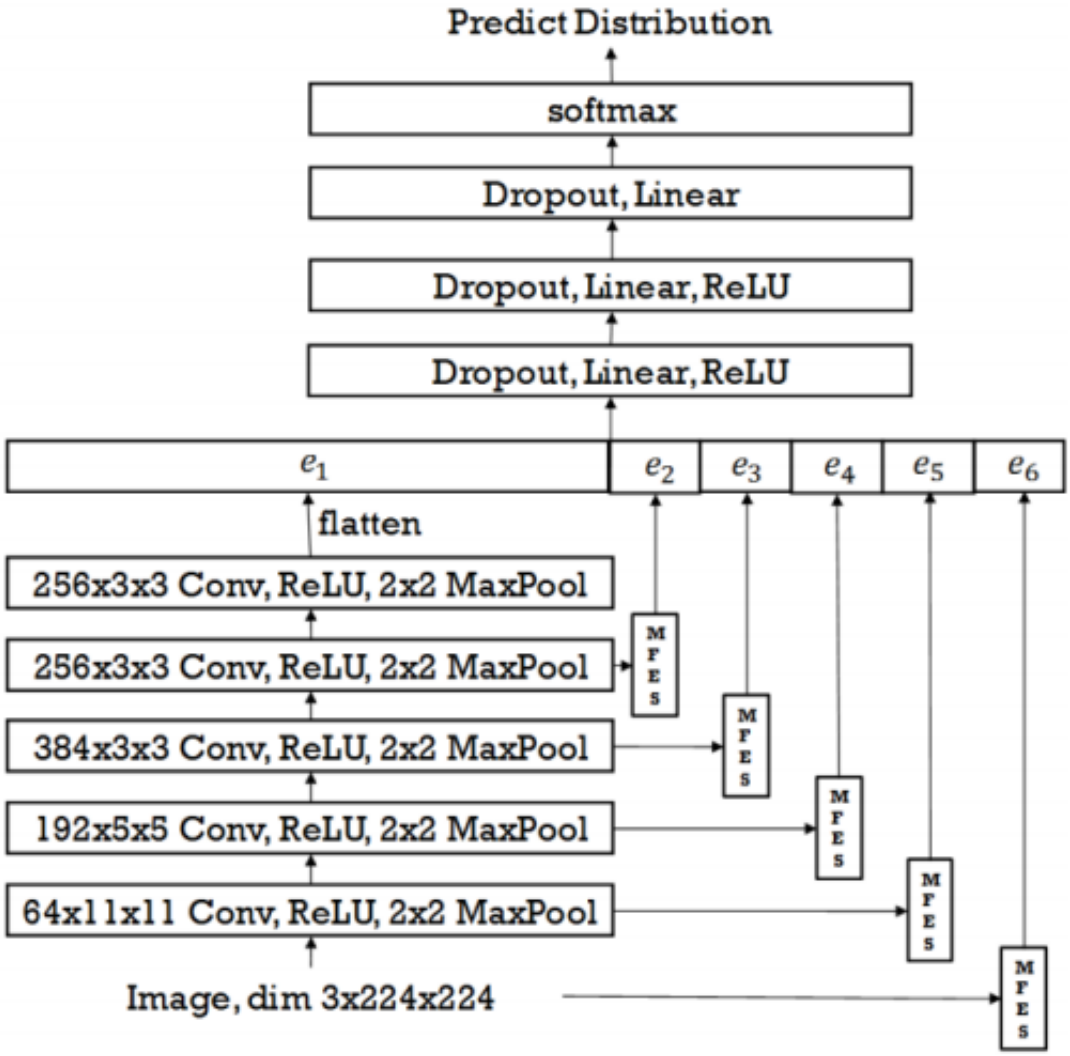


图 3-2 AlexNet-MFN 网络结构图

> MFIC：多级特征交叉分类器

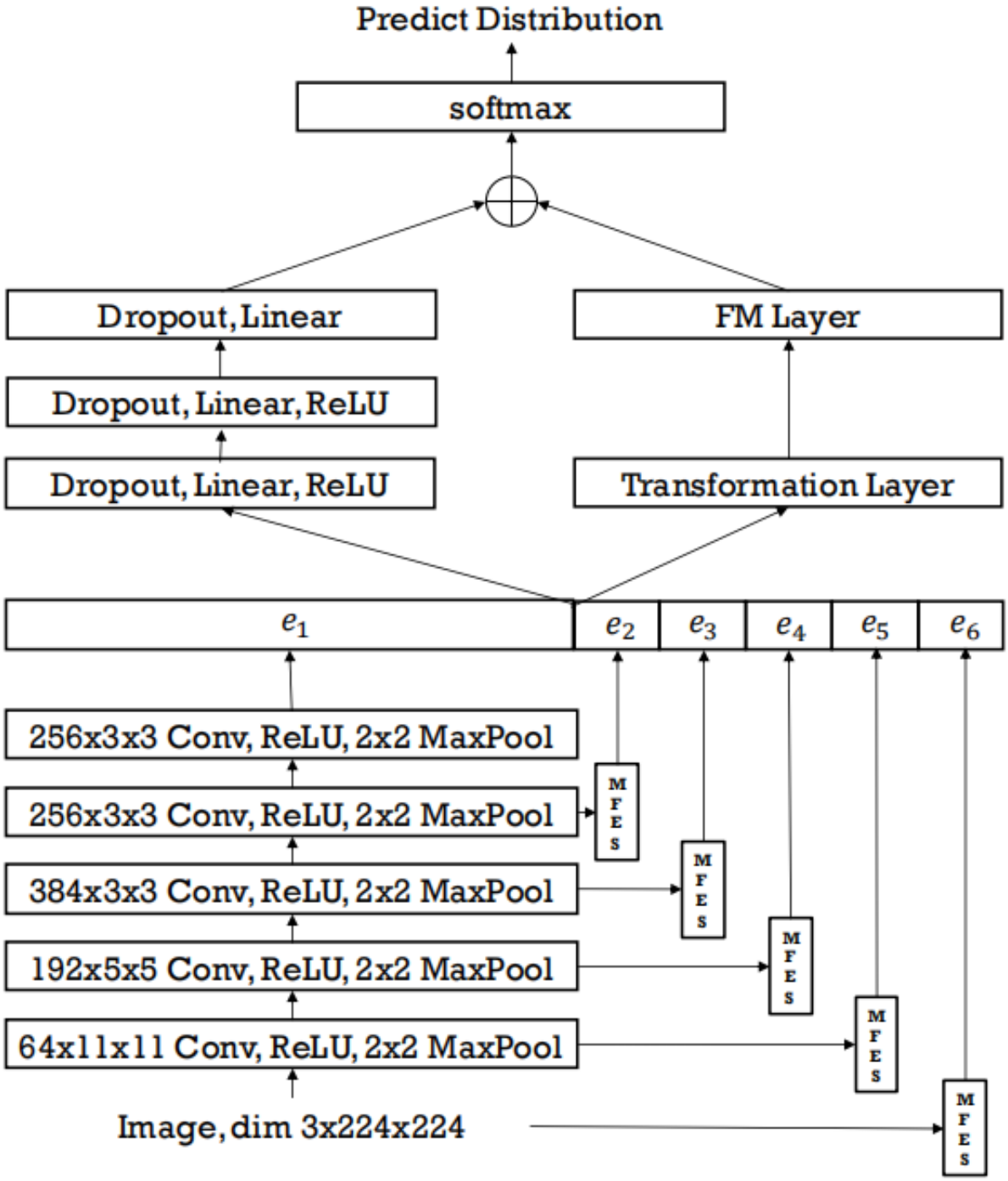


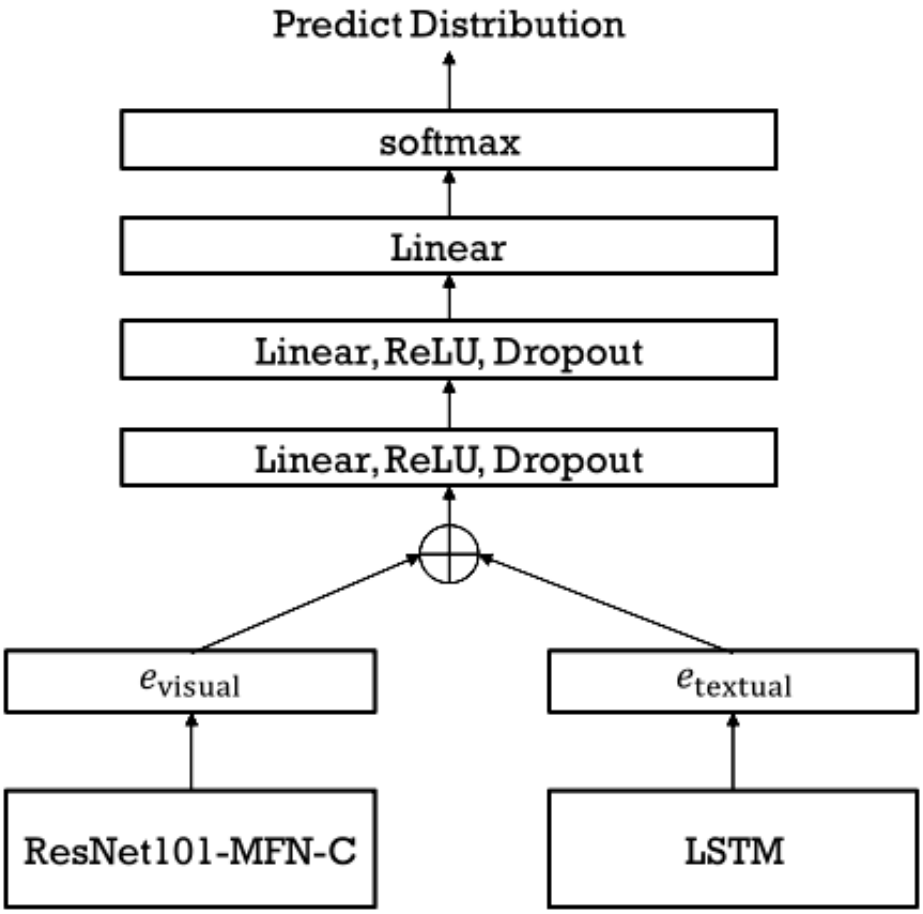
图 3-4 AlexNet-MFIC 模型结构

模型训练方式

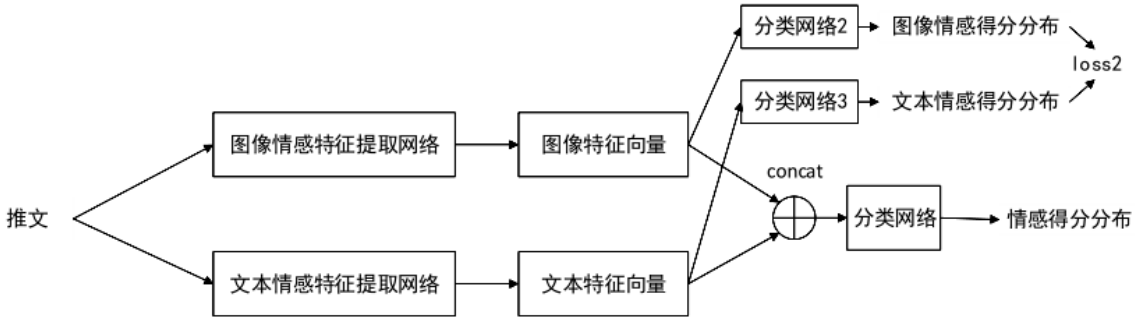
- TIS算法：自动标注算法
- CH算法：图像去重算法

图像文本联合情感分类

- 图像文本结合模型：VTFN图像文本混合网络



- 相关性损失函数



细粒度情感分析

情感要素抽取

CRFs条件随机场

CRF++主要用于标注数据和切分数据，常见的应用有词性标注、命名实体识别、文本组块分析与信息抽取等。

- 特征选取：本文选取词特征、词性特征、上下文信息特征。尽可能得标注出情感要素信息。
- 构建标注集：将标注集设定为四种简单的标记来防止特征稀疏。给定输入序列$W = \{w_i\}$，输出标注序列$Y = y_t, y_t \in \{FO, SO, ADV, P\}$

表 3-3 标注集示例

词语类别	类别说明	标注集	示例
情感对象	评价对象的属性特征	〈FO〉	外观，分辨率，性价比，配置，价格….
评价词	情感词	〈SO〉	漂亮，清晰，流畅，便宜…
	情感修饰词	〈ADV〉	很，特，太，不，没，非…
其他	情感无关词，标点符号等	〈P〉	我，到，买，，。：、；…

- 标注：输入数据为已经预处理后的意群集，通过 CRFs 进行情感要素序列化标注。

隐式情感对象识别

- 用 F 代表评论句中的情感对象，O 代表评论句中的情感词，组成特征观点对<F,O>，构建训练文档。
- 训练朴素贝叶斯文本分类器，分类识别隐式情感对象。
- 然后采用训练好的分类器完成所有预处理后语料集中评论句的情感要素识别。

细粒度情感分析

情感歧义搭配词典

- 挖掘关联规则
- 挖掘词语间搭配关系强度——点互信息PMI
- 构建成情感歧义搭配词典

正负向情感强度分析算法

情感倾向计算算法的基本思想：利用情感词和影响情感的情感修饰词计算情感对象情感。算法步骤如下：

1. 根据否定词n_i的情感值Neg_i 和程度副词d_i的情感值Mod_i ，计算情感修饰词的情感影响因子Q_{adv_i} 。
2. 结合情感词的极性P_i计算情感要素组成的属性观点对的情感极性值$Score(f_{t_i})$。
3. 计算评论语料中产品属性的正向情感强度$Sentiment(f_{t_i})_+$ 和负向情感强度$Sentiment(f_{t_i})_-$ 。

对立观点情感摘要

得到各聚合后特征观点对的正向和负向的情感强度之后，可直接由此生成情感摘要。

实训计划

内容：食品安全事件检测与舆情分析

1 / 第一周

- 余连玮：理清《网络评论文本的细粒度情感分析研究》和《文本话题识别算法的研究与实现》论文所用到的技术；

- 王子昂：理清《基于网络数据的食品安全事件检测与分析》和《社交网络推文情感分类系统的设计与实现》所用到的技术，将论文中涉及的技术整合到一个框架中，确定之后的研究和开发方向

- 两人共同完成：讨论并将上述论文的技术整合到一个框架，确定后续深入的方向。

- 输出：得到一个整合上述所有技术的大纲。

2 / 第二周

- 王子昂：配置环境，整理命名实体识别代码（标注、识别）、话题提取部分、事件检测部分的代码，理解代码结构和实现方式。基于源代码的基础上进行代码的重构，添加必要注释，修改其中存在的问题，使整体结构清晰，代码易于重用。结合得到的新的数据源，重新进行数据处理，命名实体识别、构建新的库语料、训练词向量、得到新的 LDA 模型。

- 余连玮：寻找新的适合的视频或音频数据源，进行文本提取，形成可调用的 api，集成到邓娜代码中。帮助王子昂重构代码。

- 输出：得到更加丰富的数据集、并对新的数据集进行数据处理、得到重新训练的可用模型。

3 / 第三周

- 余连玮：复现《文本话题识别算法研究与实现》中 VSM 模型部分，与 LDA 模型相结合计算相似度。

- 王子昂：复现《文本话题识别算法研究与实现》中结合 LDA、VSM 模型得到的相似度，进行文本聚类，实现话题提取。

- 输出：得到一个改进后能提取话题的系统。

4 / 第四周

- 余连玮：复现《社交网络推文情感分类系统的设计与实现》中的图像情感分类模型的搭建，包括 MFES 多级特征提取器模型，以及 MFIC 多级特征交叉分类器。

- 王子昂：复现《社交网络推文情感分类系统的设计与实现》中的图像文本联合分类算法中训练方式的实现，包括两个算法（TIS 算法：自动标注算法 CH 算法：图像去重算法）。
- 输出：实现一个图像情感分类模型。

5 / 第五周

- 共同完成：寻找中文推文情感数据集，搭建 LSTM 神经网络模型提取文本

特征向量，复现《社交网络推文情感分类系统的设计与实现》中的图像文本混合网络（VTFN），引入相关性损失函数，用于模型的训练。

- 输出：得到一个可运行的中文社交网络推文（图文）情感分类系统。

5 / 第六周

- 共同完成：将图像文本联合分类算法整合进食品安全事件的舆情分析部分，一方面将只针对微博、知乎的文本评论转换为更加丰富的图像文本联合信息，另一方面得到更细粒度的情感划分。
- 输出：得到一个完整的食品安全事件检测与分析系统：采集文本、视频、音频数据，进行事件检测与提取，最终对相关社交网络的推文进行细粒度的情感分析。