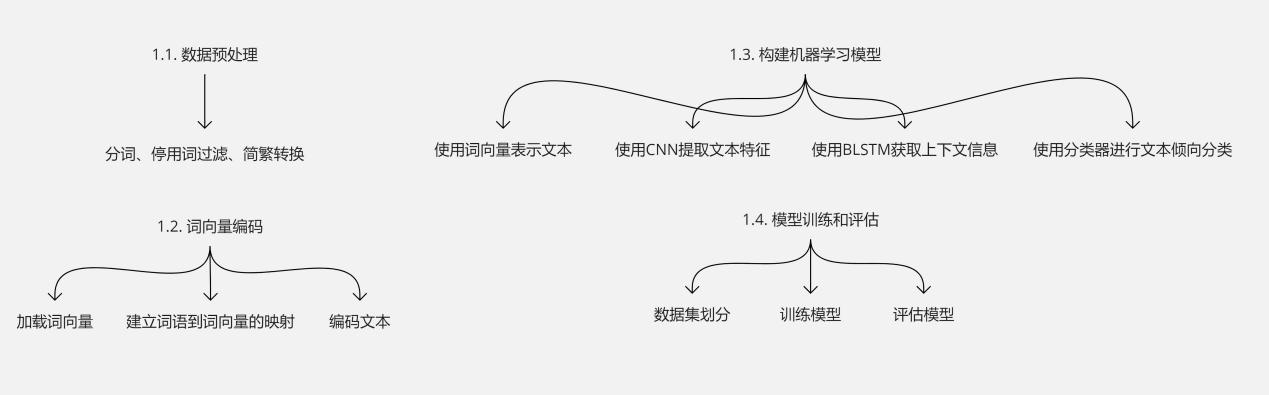
# **情感分类**

智能舆情分析应用中的情感分类模块旨在识别文本的情感倾向（正面、负面、中性），并进一步分析情感的强度。进行基于深度学习的情感分类，首先要对语句进行分词、停用词、简繁转换等预处理，然后进行词向量编码，然后利用LSTM或者GRU等RNN网络进行特征提取，最后通过全连接层和softmax输出每个分类的概率，从而得到情感分类。



* 1. **数据预处理**

**功能**：对原始数据进行处理，包括分词、停用词过滤、简繁转换等。

**关键问题**：确保文本中的噪声和无关信息最小化，提升后续模型的准确性。

**步骤：**同上文

* 1. **文本编码**

**功能**：将文本转换为向量，以便输入到深度学习模型中。

**关键问题**：选择合适的词向量库，处理特定场景下的缺失词向量。

**步骤**：

1. 加载词向量。

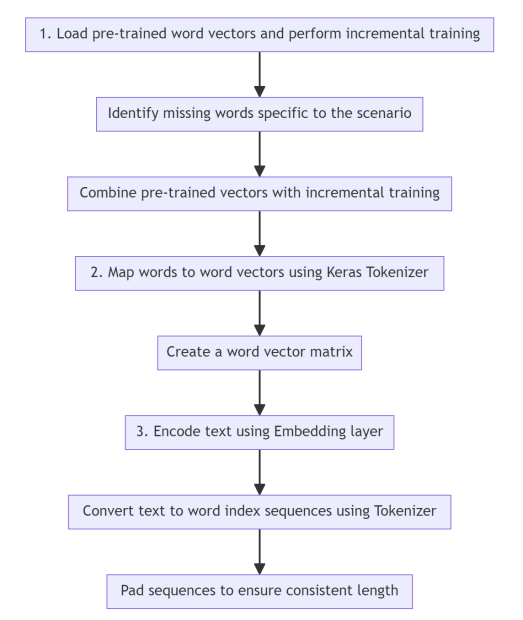
加载预训练的词向量（如Word2Vec, GloVe），并结合特定场景的增量训练。网上下载的词向量获取简单，往往缺失特定场景的词语。比如大众点评菜品场景下的鱼香肉丝、干锅花菜等词语。而自己训练则需要大量的语料，训练时间长，成本较高。所以将两种方法结合。

1. 建立词语到词向量的映射。

将文本中的词语映射到对应的词向量，并创建一个词向量矩阵以供模型使用。使用Keras的Tokenizer来建立词语索引，并创建一个词向量矩阵，词向量矩阵的每一行表示一个词语的词向量。矩阵的行数等于词汇表的大小加1（考虑到索引从1开始）。

1. 编码文本。

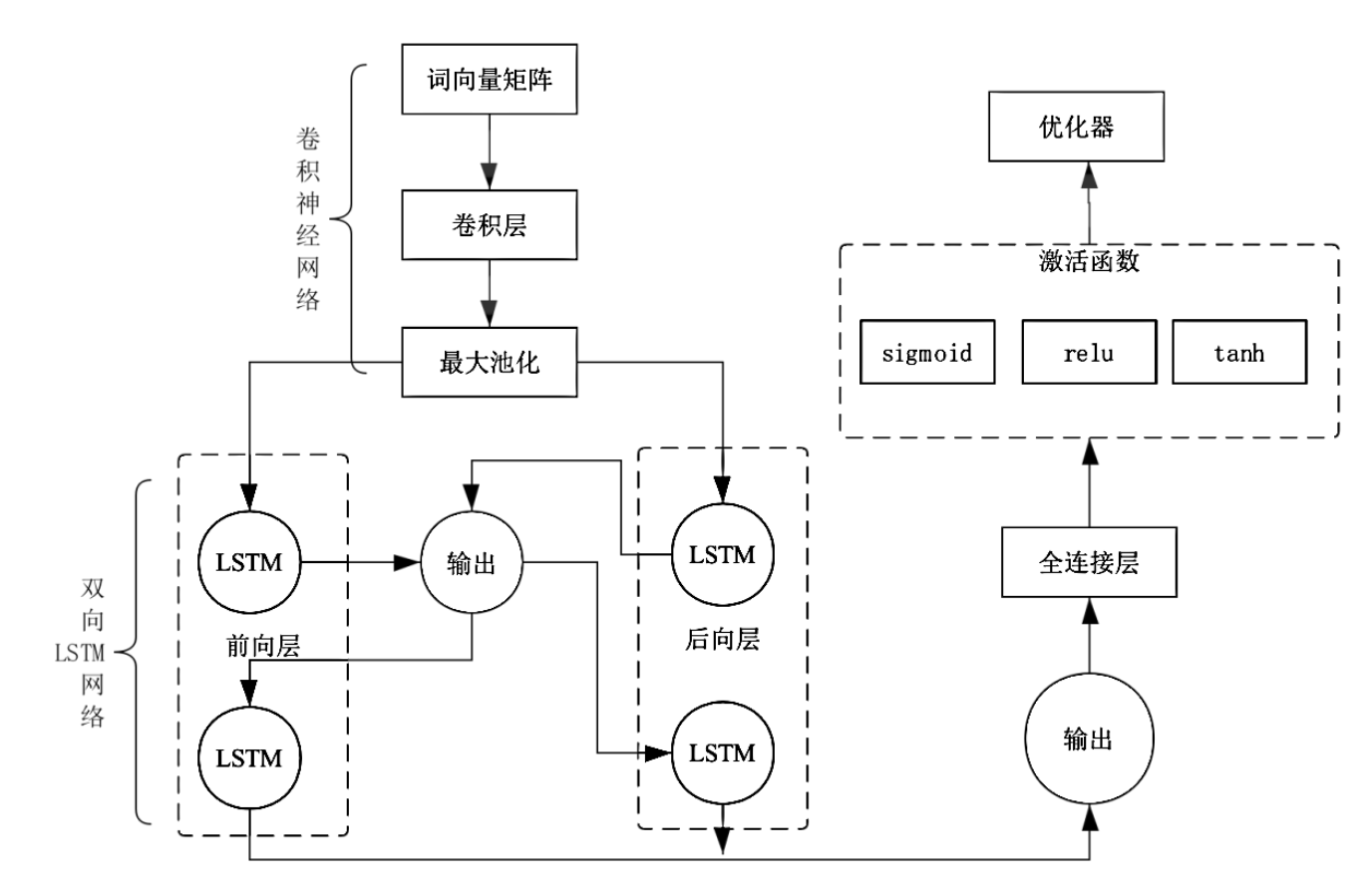
使用嵌入层将文本中的词语转换为向量表示，并对文本进行数字编码和填充。使用Tokenizer将文本转换为词语的索引序列。填充序列以确保长度一致，方便后续模型处理。使用Keras的Embedding层将词语索引转换为词向量。



* 1. **构建机器学习模型**

结合CNN和BLSTM模型来提取和分析文本特征。

CNN 的优势是可以从全局信息中提取序列特征，并考虑这些特征之间的关系，

而 BLSTM 不仅解决了长期依赖的问题，同时也能考虑上下文的关系。因此我们可以将两个模型结合起来使用进行探索，用 CNN 卷积层将提取局部特征，然后 BLSTM 层将使用特征排序来了解输入的文本排序。

**功能**：通过CNN提取局部特征，通过BLSTM获取上下文信息，最终进行分类。

**步骤：**

1. 使用词向量表示文本。

与上述步骤一致。

1. 使用 CNN 提取文本特征。

通过卷积和池化操作提取有效的特征。CNN 由输入层、卷积层、池化层和输出层构成。第一层是输入层，使用向量矩阵作为 CNN 模型的输入，就是把语料库中的矩阵作为输入进行词嵌入；第二层是卷积层，利用卷积核从字向量矩阵提取局部特征，词向量为 96 维时，采用 3×100,4×100,5×100 大小滤波器各 250 个，padding 设置为 VALID，strides 为 1；第三层是池化层，将不用的特征进行舍弃，保留有用的特征，生成固定维度的特征向量。

1. 使用 BLSTM 获取上下文信息。

BLSTM 使用前向 LSTM 获得输入序列历史信息，后向的 LSTM 可以获得输入序列的信息。两个双向 LSTM 被堆叠，第一个双向 LSTM 输出返回序列作为第二个双向 LSTM 的输入，第二个双向 LSTM 输出时前向和后向最后一个单元输出的串联，最后进行叠加。BLSTM 的第一层是嵌入层，把嵌入层的矩阵作为输入，词向量维度为 96 维；第二层和第三层都是隐藏层，隐藏层大小为250。

1. 使用分类器进行文本倾向分类。

文本倾向分类就是将全连接层的输出输入到激活函数层进行预测，如果预测结果是 Y，实际上是 Y’，则使用交叉熵作为损失函数，然后利用优化器对模型参数进行优化，同样的为了防止过度拟合，使用 dropout 机制，dropout 概率为 20%。

* 1. **模型训练和评估**

功能：对模型进行训练，并使用验证集评估模型性能。

关键问题：数据集划分、训练过程中的过拟合问题。

步骤：

1、数据集划分：将数据集划分为训练集和验证集。

2、训练模型：使用训练集对模型进行训练。

3、评估模型：使用验证集评估模型性能，并调整参数。