**自然语言处理大作业报告**

**—— 主题层次的情感分类**

# 任务及数据集介绍

该项目作业的具体任务是来自于BDCI2018-汽车行业用户观点主题及情感识别的题目。数据是网络中公开的用户对汽车相关内容的评价文本。此任务是对每条文本内容（即用户评论）进行分析，确定该条评论中讨论的主题以及每个主题的情感信息，以此获知用户对所讨论主题的偏好。

其中主题一共分为10类：动力、价格、内饰、配置、安全性、外观、操控、油耗、空间、舒适性。情感分为3类：中立、正向、负向，分别用数字0、1、-1表示。

训练集数据一共有10654条评论，已由人工进行标注，为CSV格式的文件，以逗号分隔，首行为表头，字段说明如表1所示。其中content\_id与content一一对应，但同一条content中可能包含多个主题，此时会出现多条记录来标注不同的主题及情感，因此在整个训练集中content\_id存在重复值。

测试数据也为CSV格式的文件，首行为表头，字段有content\_id和content，与训练集对应字段相同。

但由于BDCI官网上此比赛已结束，线上提交接口已关闭，我们无法在测试数据对模型进行评估。因此，此次项目作业完全是在训练数据上进行的。我们将训练数据按照8:1:1的比例划分成训练集（8523条数据）、验证集（1066条数据）、测试集（1065条数据）。

表1： 训练集字段说明

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 类型 | 描述 | 说明 |
| content\_id | Int | 数据ID |  |
| content | String | 文本内容 |  |
| subject | String | 主题 | 分析出的主题 |
| sentiment\_value | Int | 情感分析 | 分析出的该主题的情感 |

该任务采用F1值评估指标，按照“主题+情感词”精确匹配的方式，用Tp表示判断正确的数量，Fp表示判断错误或多判的数量，Fn表示漏判的数量。则准确率 ，召回率 ，。

# 相关工作

上述任务是一个主题级别（Aspect-level）的文本情感分类任务，它是情感分类的一个分支，其目标是识别出句子中某个特定主题的情感极性。相比文档级别的情感分类任务，主题级别的情感分类是更细粒度的分类任务，它可以克服文档级别情感分类的局限性。尤其是当一篇文本中同时涉及多个主题时，正负的情感极性通常是混杂在一起的，如果忽略主题信息，我们很难确定目标文本的情感极性。比如“great food but the service was dreadful”，在这条用户对餐厅的评论中同时涉及到了“食物”和“服务”两个主题，而且用户在这两个主题上的评价的情感极性是相反的，如果我们在文本层次进行情感分类，很难说这条文本是正向的，还是负向的，甚至是中性的。

在主题级别的情感分类任务上，目前已有的工作大致可以分为两类。一类是基于手工提取的特征和机器学习算法的级联方法，例如Nasukawa等人[1]的工作中，首先对句子进行依存句法分析，然后使用预定义的规则确定每个主题的情感极性。Jiang等人[2]提出，基于句子的语法结构构造一些主题依赖的特征，将这些特征和其他文本内容特征一起作为SVM分类器的输入，从而提高主题依赖的情感分类的准确度。另一类工作是采用端到端的神经网络模型，使用神经网络模型可以自动进行表征学习，不需要花费大量时间进行手工工程或者编写规则。在这类模型中，大多数是基于LSTM的网络架构，因为LSTM可以很好地处理像文本这样的序列数据，而且可以捕捉到长期依赖。例如Tang等人[3]提出了TD-LSTM的方法，该方法的基本思路是通过两个LSTM网络分别来建模主题的左、右上下文，然后使用这两个LSTM网络的输出的串接进行情感预测。除此之外，近年来的工作中也大量使用注意力机制，它和LSTM结合，在很多任务上可以提高网络性能。例如Wang等人[4]提出了AE-LSTM模型和ATAE-LSTM模型。AE-LSTM模型是先将文本通过LSTM网络层，得到每个词对应的隐单元输出，然后结合主题向量表示计算一组注意力值来加权句子的不同部分，从而得到最终的向量表达，进行情感分类。相比AE-LSTM模型，ATAE-LSTM模型更强调主题向量表达的影响，它把主题向量表达和每个词的向量表达串接在一起进行建模，得到文本的表示。Ma等人[5]提出了IAN模型，该模型使用两个LSTM网络分别对句子和主题进行建模，然后对句子的隐状态进行池化操作，使用池化的结果和主题的隐状态计算一组注意力值，对主题的隐状态进行加权得到主题的最终向量表示，同时对主题的隐状态进行池化操作，使用池化结果和句子的隐状态计算另一组注意力值，对句子的隐状态进行加权得到句子的最终向量表示，最后将两个向量表示串接起来进行情感分类。

# 模型

针对BDCI-2018的汽车行业用户观点主题及情感识别任务，我们在调研的基础上对一些模型进行改进，从而使其能够适应该任务的特点和需求。最终实现的3个模型的具体思路如下。

## 3.1 基于AOA的改进模型1

第一个模型架构如图1所示，是对AOA（Attention Over Attention）模型[6]的改进，一共分为四个部分：词嵌入模块、双向LSTM模块、注意力模块和最终的分类预测模块。

**词嵌入模块**：对于一个给定的句子，经过分词之后可以将其表示为。对于所有主题，我们可以将其表示成 （此任务有10个主题词，则m=10）。我们首先使用预训练好的词向量将每个词映射到一个低维的实值向量。则句子和主题分别通过一个词嵌入层之后我们可以得到两个词向量集合 和 。

**双向LSTM模块**：我们将得到的词向量s和t分别通过一个双向LSTM层来学习句子和主题的隐含语义表达，结果为双向LSTM隐单元输出的串接，此时可以得到句子和主题的表达 和 。

**注意力模块**：对于上一步得到的 和 ，我们首先计算一个相互作用矩阵 ，矩阵中的每个元素 表示句子中的某个词 和某个主题 之间的关系。然后我们对矩阵 的每列进行softmax操作，得到aspect-to-sentence的注意力值 ，其中 。直观来讲， 每一列的意义是，对于某个主题而言，句子中的哪些词比较重要。最终根据注意力值 和句子的隐含语义表达 来计算各主题感知的句子向量表达，即 。

**分类模块**：我们将上一步得到的句子表达向量 展平，然后输入到两层的全连接网络，得到40维的分类结果。对于10个主题中的每一个主题，对应有4维输出结果，分别表示句子中不含该主题的概率、该句子在该主题的情感极性为中性的概率、该句子在该主题的情感极性为正向的概率、该句子在该主题的情感极性为负向的概率。对每个主题的四维输出结果进行softmax，以此判断该句子是否不含该主题或者该句子在该主题上的情感极性。

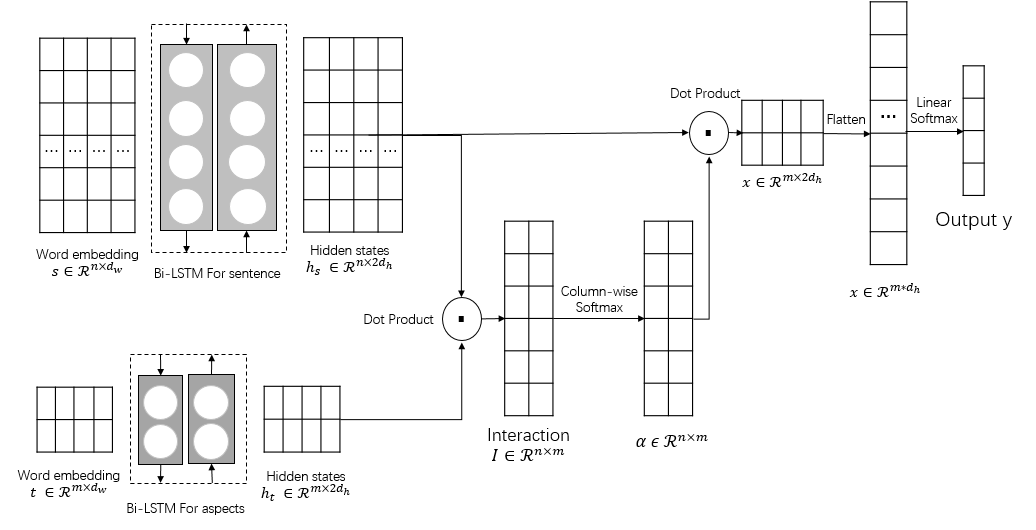


图1：AOA模型整体架构

## 3.2 基于IAN的改进模型2

第二个模型架构如图2所示，是对IAN（Interactive Attention Networks）模型的改进，一共分为四个部分：词嵌入模块、双向LSTM模块、交互注意力模块和最终的分类预测模块。

**词嵌入模块**：与上一个模型相同，句子和主题分别通过一个词嵌入层之后我们可以得到两个词向量集合 和。

**双向LSTM模块**：将得到的词向量s和t分别通过一个双向LSTM层来学习句子和主题的隐含语义表达，结果为双向LSTM隐单元输出的串接，此时得到隐状态 和 。

**交互注意力模块**：对于上一步得到的 和 ，我们首先进行平均池化，得到 和 。 然后由和句子的隐状态计算一组注意力分值，其中每个分值，由注意力值对句子的各部分进行加权，得到句子的最终向量表达。同时由和主题的隐状态计算另一组注意力分值，其中每个分值，由该组注意力值对各主题进行加权，得到主题的最终向量表达。两组注意力分值的计算公式中，。

**分类模块：**我们将上一步得到的主题和句子的最终向量表达和串接起来，然后输入到两层的全连接网络，得到40维的分类结果。对于10个主题中的每一个，有4维输出结果，分别表示句子中不含该主题的概率、该句子在该主题的情感极性为中性的概率、该句子在该主题的情感极性为正向的概率、该句子在该主题的情感极性为负向的概率。这一步与上一个模型相同，对每个主题的四维输出结果进行softmax，以此判断该句子是否不含该主题或者该句子在该主题上的情感极性。

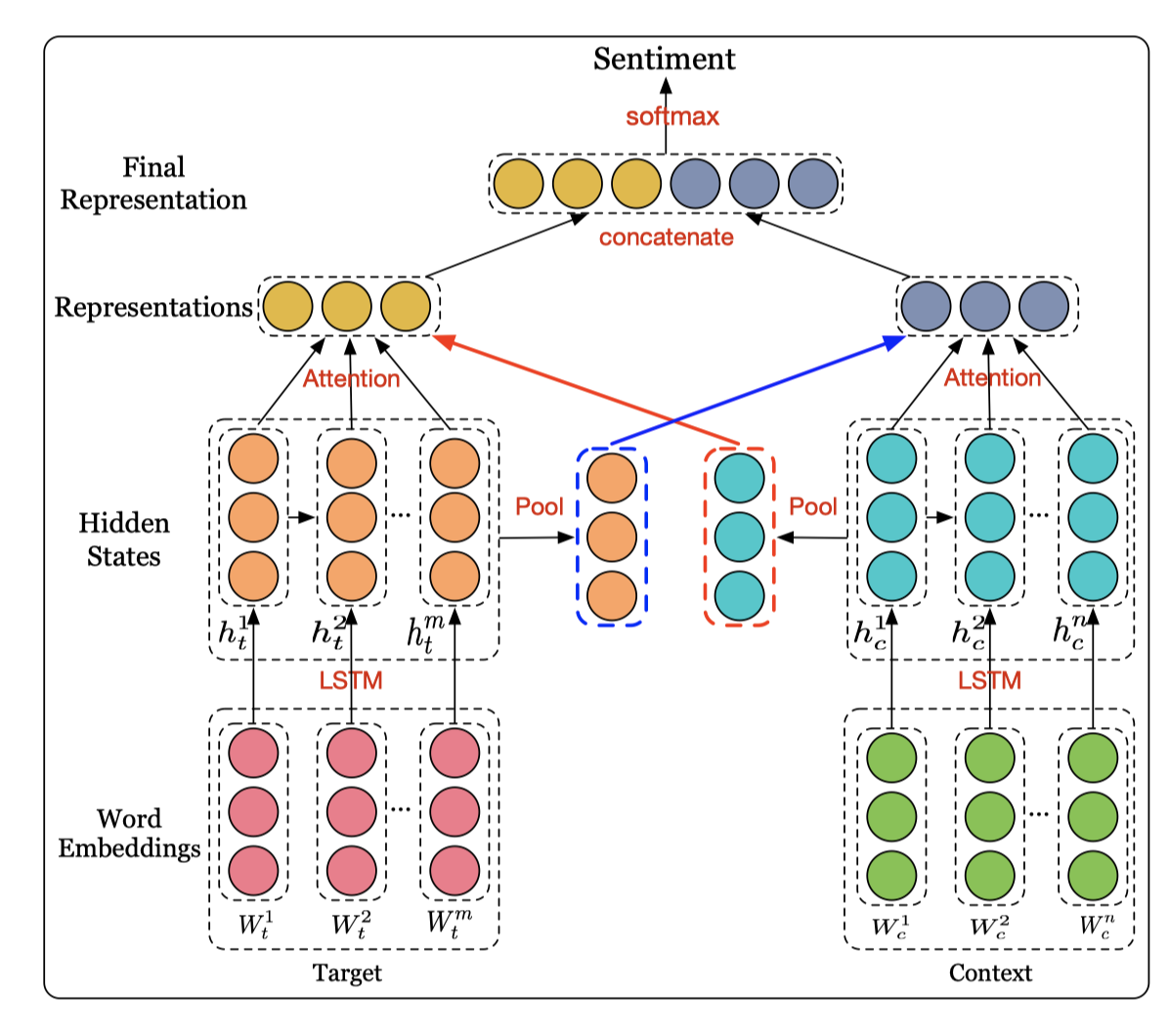


图2：IAN模型整体架构

## 3.3 基于AEN的改进模型3

第三个模型架构如图3所示，是利用AEN（Attention Encoder Networks）模型[7]对AOA（Attention Over Attention）模型的更深层次改进，一共分为五个部分：词嵌入模块、双向LSTM模块、注意力模块、多输入池化模块和最终的分类预测模块。

**词嵌入模块**、**双向LSTM模块**、**注意力模块**：这三个模块与第一个模型AOA相同，总共得到三个输出向量，句子的隐层状态 和主题的隐层状态 以及句子和主题交互得到的主题交互状态 ，这三个向量都是下一个模块的输入。

**多输入池化模块**：将上一步得到的三个输出分别进行一维最大池化，得到三个池化结果向量，对应为、以及，然后将三个池化向量串接，得到最终分类模块的输入向量x 。

**分类模块**：最终的分类模块输入为上一步三个池化向量的串接展平结果，分类机制与前两个模型相同。对每个主题的四维输出结果进行softmax，以此判断该句子是否不含该主题或者该句子在该主题上的情感极性。

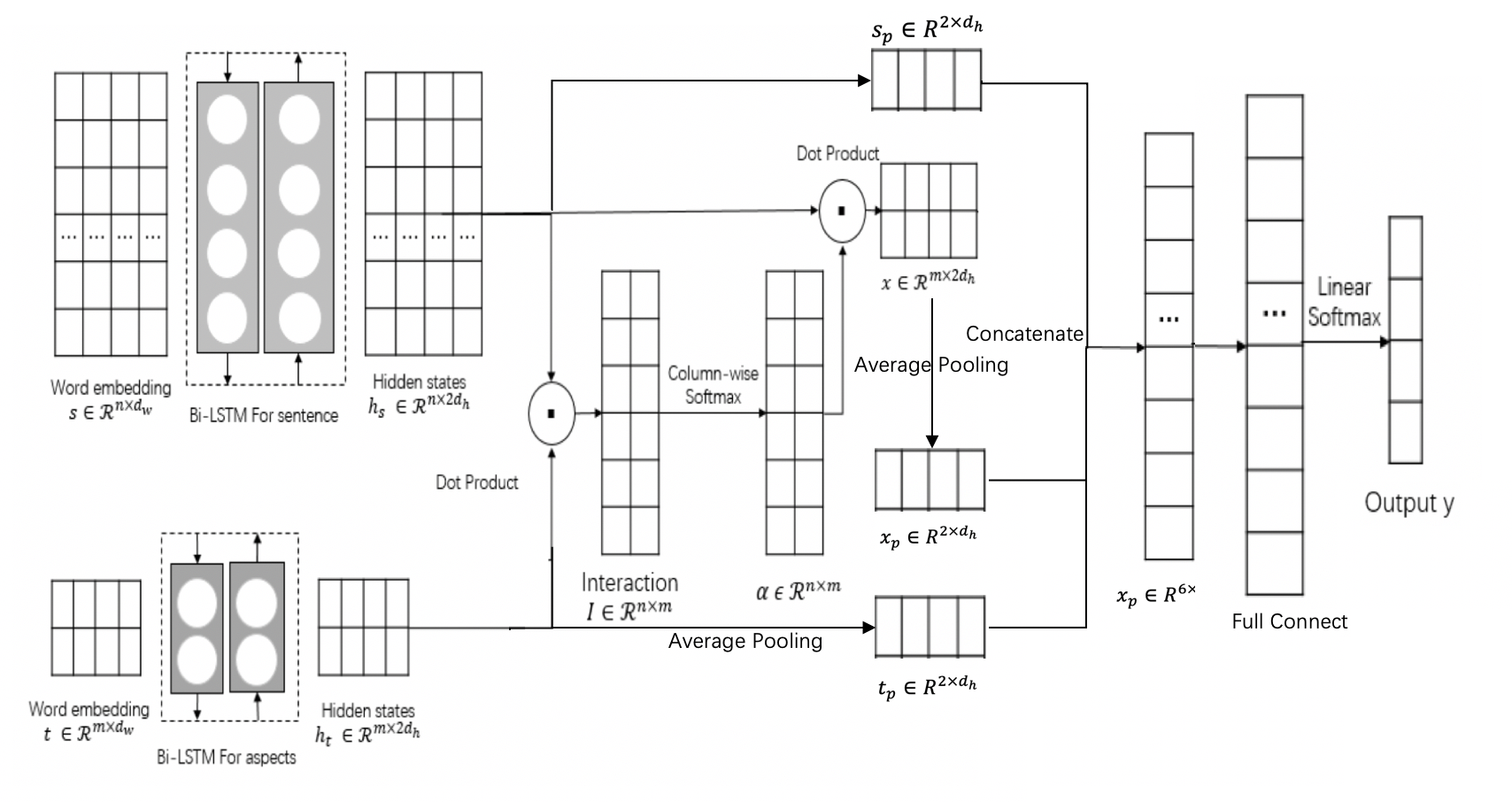


图3：AEN模型整体架构

# 实验过程与结果

## 4.1 实验过程

第一步：对任务数据进行分词，得到数据集的词典，及每个词的embedding，结果文件embeddings\_.p会输出到指定目录下。（因为本实验使用腾讯公开的预训练的汉语词嵌入，数据量比较大，为了减少模型训练时的计算资源需求，我们先处理得到本任务数据集的词典，并提取出这些单词的embedding另存到文件embeddings\_.p）

|  |
| --- |
| python get\_embed.py \  --train\_data\_path=path/to/train\_data \  --embed\_data\_path=path/to/pretrain/word\_embedding \  --out\_embed\_dit=dir/to/save/output\_file |

第二步：模型训练及评估。如前所述，我们将训练集以8:1:1的比例划分成训练集、验证集和开发集进行实验，此处我们的超参数具体设置如下所示。

|  |
| --- |
| python aspect\_senti.py \  --train\_data\_path=path/to/train\_data \  --embed\_data\_path=path/to/embeddings\_.p \  --out\_dir=/dir/to/save/model \  --max\_sentence\_length=256 \  --lstm\_hidden\_size=100 \  --dense\_hidden\_size=512 \  --leakyRelu\_alpha=0.01 \  --drop\_rate=0.5 \  --reg\_rate=0.001 \  --optimizer=adam \  --lr\_factor=0.65 \  --batch\_size=32 \  --epochs=30 |

为了方便运行，模型必须的参数都已经在代码中设置了默认值，均是在验证集上进行调参之后的最佳参数设置，而且项目目录已经按照需要进行调整，只需输入python加模型名称即可运行不同的模型，比如python aen.py。

## 4.2 实验结果

三个模型训练时使用的损失函数均是各主题的交叉熵损失之和。训练过程中我们保存在验证集上值最大的模型，然后在测试集上进行评估，同时每个模型还测试了不同的固定文本长度。三个模型在测试集上的评估结果如下表：

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Text\_length | Accuracy | F1 |
| AOA | 60 | 0.943 | 0.724 |
| 256 | 0.943 | 0.723 |
| IAN | 60 | 0.941 | 0.705 |
| 256 | 0.938 | 0.688 |
| AEN | 60 | 0.943 | 0.720 |
| 256 | 0.943 | 0.722 |

据BDCI官网上该任务的排行榜，官方测试集上最好的模型表现为F1=0.708。由于我们无法再官方测试集上进行模型评估，因此我们的结果和排行榜结果无法进行的严格比较，但从结果来看，我们的模型还是表现不错的。

# 总结

通过此次项目实践，我们了解了文本情感分析的相关模型，尤其是主题级别的情感极性分析，改进了端到端的神经网络模型，如AOA、IAN和AEN，使得模型能够适应主题层次情感分类任务的特点和需求。我们遇到的问题要比普通的文本分类任务更加复杂，在该问题中，一篇文本中通常会同时涉及多个主题，正负的情感极性通常是混杂在一起的，如果忽略主题信息，我们很难确定目标文本的情感极性。因此，在在主题级别的情感分类任务上，我们在调研的基础上对AOA、IAN和AEN模型进行了改进，从而使其能够适应该任务的特点和需求。

在BDCI官方数据集上，我们将训练集以8:1:1的比例划分成训练集、验证集和开发集进行实验，设置模型的固定文本长度分别为60和256，三个模型的表现不错，其中AOA模型和AEN模型的F1值均超过了官网测试集上最好的模型表现。由于我们无法对官方发布的测试集进行线上评估，因此无法和赛题排行榜进行精确对比，但就排行榜情况来看，最好成绩也不高于0.71，我们的方法所取得的结果是不错的。

情感分析是自然语言处理中常见的场景，比如汽车行业用户观点主题及情感识别，对于指导产品更新迭代具有关键性作用。通过对用户评价进行情感分析，可以挖掘产品在各个维度的优劣，从而明确如何改进产品。比如对用户评价进行分析，可以得到汽车动力、价格、内饰、配置、安全性、外观、操控、油耗、空间、舒适性等多个维度的用户情感指数，从而从各个维度上改进现在汽车行业的制造和销售状况。

深度学习在NLP领域内应用相当广泛，在情感分类领域，我们同样可以采用深度学习方法。基于深度学习的情感分类，具有精度高，通用性强等优点。通过完成这次项目作业，我们对情感分类问题有了更深入的认识，对神经网络模型有了进一步的掌握，获益良多。

参考文献：

1. Nasukawa, T., Yi, J.: Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2nd international conference on Knowledge capture. pp. 70–77. ACM (2003)

2. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 151–160. Association for Computational Linguistics (2011)

3. Tang, D., Qin, B., Feng, X., Liu, T.: Effective lstms for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 3298–3307 (2016)

4. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based lstm for aspect-level sentiment classification. In: EMNLP. pp. 606–615 (2016)

5. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 4068–4074 (2017)

6. Huang B, Ou Y, Carley K M. Aspect level sentiment classification with attention-over-attention neural networks[C]//International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, Cham, 2018: 197-206.

7. Song Y, Wang J, Jiang T, et al. Attentional Encoder Network for Targeted Sentiment Classification[J]. arXiv preprint arXiv:1902.09314, 2019.