Bilibili 中文弹幕情感分析技术路线

一、情感词典

1、技术原理

基于情感词典的方法将文本情感看作词汇级情感的线性叠加，核心步骤如下：

1) **分词与词性标注**：对弹幕 D 进行分词，得到词序列 $W = \{w_1, w_2, \ldots, w_n\}$。

2) **情感词匹配**：将 W 与情感词典 $\mathcal{L}$ 比对，提取情感词子集 $W_{\text{emo}} \subseteq W$，并获取对应的情感极性 $s_i \in \{-1, 0, +1\}$（负、中、正）及强度 $\alpha_i \geq 0$。

3) **情感得分聚合**：常用线性加权求和模型

$$\text{Score}(D) = \sum_{w_i \in W_{\text{emo}}} \alpha_i s_i \tag{1}$$

若 $\text{Score}(D) > 0$ 则判为正向，$\text{Score}(D) < 0$ 为负向，否则为中性。

4)（**可选**）**强度修正**：考虑否定词、程度副词。设否定词集合 $\mathcal{N}$，程度副词集合 $\mathcal{D}$，强度系数 $\beta_j \in [0.5, 2.0]$，则修正后得分

$$\text{Score}_{\text{rev}}(D) = \sum_{w_i \in W_{\text{emo}}} \alpha_i s_i \cdot \prod_{w_k \in \text{window}(w_i)} (-1)^{\mathbf{1}_{w_k \in \mathcal{N}}} \cdot \prod_{w_m \in \text{window}(w_i)} \beta_m^{\mathbf{1}_{w_m \in \mathcal{D}}}. \tag{2}$$

2 优点

- **实现简单，无需训练数据**：王春东等指出“计算简单且不需要额外的资源”，对硬件零要求，可在个人笔记本上秒级完成数万条弹幕打分。
- **可解释性强**：每个得分均可回溯到具体情感词，便于舆情干预时定位敏感词。

3 缺点

- **词典构建耗时且领域迁移困难**：文献强调“情感词典的构建是一个工作量巨大且复杂的工程”“领域内适用，通用性较差，鲁棒性较低”。弹幕不断产生新词（如“awsl”“蚌埠住了”），若词典更新滞后，极易误判。
- **对中文修辞敏感性差**：中文存在“一词多义、反讽”现象，同一词汇在不同语境下情感极性可能反转；基于词典的线性模型难以利用上下文信息，导致精度低于机器学习与深度学习方法。
- **时效性不足**：弹幕 (类似微博)“每天都会产生大量网络新词，若情感词典不能及时更新，就很有可能误判”。

二、传统机器学习

1. 原理概述

传统机器学习把情感分析视为有监督文本二分类问题，核心流程如下：

1) **语料准备**：对弹幕文本集 $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ 进行人工标注，$y_i \in \{+1, -1\}$ 表示正向/负向。

2) **文本向量化**：采用词袋模型，先分词得到词汇表 V，再计算每句的 TF-IDF 权重

$$\text{tf-idf}(t, d) = \frac{f_{t,d}}{\sum_{k \in d} f_{k,d}} \cdot \log \frac{N}{n_t}, \tag{3}$$

其中 $f_{t,d}$ 为词 t 在弹幕 d 中的出现次数，n_t 是含 t 的弹幕数，N 为总弹幕数。最终每条弹幕被表示为稀疏向量 $\mathbf{x}_i \in \mathbf{text}R^{|V|}$。

3) **特征选择**：用卡方检验或信息增益筛选高区分度特征，降低维度。

4) **分类器训练**：常用算法

- 朴素贝叶斯（NB）：假设特征条件独立，最大化后验概率

$$\hat{y} = \arg \max_{y \in \{+1, -1\}} P(y) \prod_{j=1}^{|V|} P(x_j \mid y). \tag{4}$$

- 支持向量机（SVM）：求解最大间隔分离超平面

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i, \quad \text{s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0. \tag{5}$$

5) **预测与评估**：在测试集上计算准确率、Precision、Recall、F1。

2. 优点

- **特征可解释性强**：权重向量 $\mathbf{w}$ 可直接反映哪些词汇对正向或负向贡献最大，便于人工审核与舆情干预。
- **小数据友好**：在仅标注数千条弹幕的情况下，TF-IDF+SVM 即可达到约 80% 的宏 F1，无需昂贵算力。
- **训练速度快**：稀疏表示下维度虽高，但非零元素少，线性 SVM 的凸优化可在 CPU 秒级收敛。

3. 缺点

- **特征工程繁重**：需要手工设计分词、停用词、n-gram、否定词窗口等，特征优劣直接决定上限；文献指出“人工特征提取的好坏直接影响分类结果”。

- **上下文建模能力弱**：TF-IDF 忽略词序与远距离依赖，无法区分“不是棒”与“真棒”，对反讽、梗、网络新词容错低。
- **领域迁移代价高**：弹幕语料时效性强，当新梗爆发时需重新标注、重新做特征选择，否则性能骤降；文献强调“中文文本上下文关联性较大”，传统方法难以捕获。

三、深度学习轻量级 BERT

1. 原理概述

轻量级 BERT 路线以预训练 + 微调范式完成弹幕情感分类，步骤如下：

1) **数据准备**：收集并标注弹幕对 (x_i, y_i)，$y_i \in \{+1, -1\}$。

2) **预训练模型**：采用中文 RoBERTa-wwm-ext（参数量 110 M，层数 12，隐层 768）。

3) **输入表示**：每条弹幕 x 被表示为

$$\text{Input} = [\text{CLS}]\, t_1 t_2 \ldots t_L\, [\text{SEP}] \tag{6}$$

其中 t_j 为 WordPiece 子词，$L \le 512$。

4) **双向上下文编码**：通过多层 Transformer 块计算上下文向量

$$\mathbf{H} = \text{Transformer}(\mathbf{E} + \mathbf{P}), \quad \mathbf{h}_{\text{CLS}} \in \mathbf{R}^{768} \tag{7}$$

$\mathbf{E}$ 为词嵌入，$\mathbf{P}$ 为位置嵌入。

5) **情感分类头**：在 $\mathbf{h}_{\text{CLS}}$ 后接 Dense+Softmax

$$P(y \mid x) = \text{softmax}(\mathbf{W}\mathbf{h}_{\text{CLS}} + \mathbf{b}), \quad \mathbf{W} \in \mathbf{R}^{2\times 768}. \tag{8}$$

6) **微调**：冻结底层或不冻结，以交叉熵损失最小化进行 2-3 epoch 微调，学习率通常设 2×10^{-5}。

2. 优点

- **无需人工特征**：模型自动提取 n-gram、远距离依赖及反讽信号，文献指出“能充分考虑到文本的上下文语义信息”。
- **精度高**：在同等 1-2 k 标注弹幕下，轻量 BERT 的 F1 通常比 TF-IDF+SVM 高 5-8 个百分点。
- **跨域迁移快**：只需替换顶层分类器并再微调几轮即可适应新视频主题，无需重新设计词典或特征。

3. 缺点

- **计算资源需求大**：110 M 参数、2 GB 显存起步，CPU 推理 100 条/秒级，远低于词典或线性模型；文献提到“梯度消失、并行计算难”等传统 RNN 痛点虽被 Transformer 缓解，但参数量仍远超传统方法。
- **可解释性差**：深层自注意力权重难以直接映射到人类可理解的“关键词”，不利于舆情审核时给出具体理由。
- **对短文本过敏感**：弹幕长度极短且噪声大，若标注样本不足易过拟合；文献亦指出“仅对单条文本编码，未考虑全局情感分布”时，模型可能放大个别极端用词。

参考文献

[1] 王春东, 张卉, 莫秀良, 等. 微博情感分析综述 [J]. 计算机工程与科学, 2022, 44(1): 165–175.